UNIVERSITY *of* TASMANIA

School of Physical Sciences

Stats, Drugs and Rock and Roll: Statistical applications to temporally autocorrelated substance use data

Melissa Ann Humphries
BScHons

November 2017

Supervisors: A. Prof. Barbara Holland
A. Prof. Raimondo Bruno

Submitted in fulfilment of the requirements for the degree
Doctor of Philosophy, Science

Declaration

This thesis contains no material which has been accepted for degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by any other person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Melissa Humphries

Statement of Published Work

Statement of Co-Authorship

The following people and institutions contributed to the publication of work undertaken as part of this thesis:

Melissa Humphries, School of Physical Sciences = **Candidate**

Raimondo Bruno, University of Tasmania = **Author 1**

Foon Lai, The University of Queensland = **Author 2**

Phong Thai, The University of Queensland = **Author 3**

Barbara Holland, University of Tasmania = **Author 4**

Jake O'Brien, The University of Queensland = **Author 5**

Christoph Ort, Swiss Federal Institute of Aquatic Science and Technology= **Author 6**

Jochen Mueller, The University of Queensland = **Author 7**

Yuliya Karpievitch, University of Tasmania = **Author 8**

Simon Wotherspoon, Institute of Marine and Antarctic Studies = **Author 9**

Author details and their roles:

Paper 1

Humphries, M. A., Bruno, R., Lai, F. Y., Thai, P. K., Holland, B. R., O'Brien, J. W., Ort, C., and Mueller, J. F. (2016). Evaluation of monitoring schemes for wastewater-based epidemiology to identify drug use trends using cocaine, methamphetamine, MDMA and methadone. *Environmental Science & Technology*, 50 (9), 4760-4768

Located in chapter 2 (Supplementary Material incorporated)

Candidate was the primary author and with authors 1, 4, 6 and 7 contributing to the idea, its formalisation and development.

The data was provided by authors 2, 3, 5 and 7 who both collected the samples and completed the chemical decomposition.

All statistical analyses were completed by the candidate with guidance and suggestions from authors 4 and 6.

Authors 2 and 3 provided the background information for the Experimental Methods.

Authors 1 through 7 assisted with refinement and presentation

Paper 2

Humphries, M. A., Bruno, R., Karpievitch, Y., and Wotherspoon, S. (2015). The expectancy valence model of the Iowa Gambling Task: Can it produce reliable estimates for individuals? *Journal of Mathematical Psychology*, 64 - 65, 17 – 34.

Located in chapter 4 (Supplementary Material partially incorporated and in chapter 5)

Candidate was the primary author and with authors 1 and 9 contributing to the idea, its formalisation and development.

The data was provided by author 1 (collected by Jessica Hartley and Elizabeth Murray).

All statistical analyses were completed by the candidate with guidance and suggestions from author 9.

Authors 1, 8 and 9 assisted with refinement and presentation

We the undersigned agree with the above stated "proportion of work undertaken" for each of the above published (or submitted) peer-reviewed manuscripts contributing to this thesis:

Signed: _____	_____
Barbara Holland	*John Dickey*
Primary Supervisor	*Head of School*
School of Physical Sciences	*School of Physical Sciences*
University of Tasmania	University of Tasmania
Date: 1/11/17	1/11/17

Abstract

The use of illicit drugs is an area of interest across a broad range of industries and fields including public policy, law enforcement and physical and mental health. Both the field of measuring drug use in the community and understanding its cognitive impacts are therefore the subject of constant research, development and innovation. This thesis examines statistical applications to temporally autocorrelated data in both the areas of drug use extent and cognitive impact. Specifically, sampling strategies for ascertaining drug use extent from waste water through the utilisation of patterns in weekly drug use is examined. This leads to a practical example of when representative sampling is cost effective enough to be a viable alternative to random sampling. The ability to ascertain cognitive impacts of drug use at the level of the individual is then explored. Analyses of empirical models of cognitive behaviour, that have been traditionally utilised to decompose behaviour, on psychological assessment tools the Iowa Gambling Task and the Balloon Analogue Risk Task are considered. We show that empirical models can lead to an inability to uniquely describe behaviour when considering individuals with possible cognitive impairments (those with extreme behaviours) and discuss the possibility of utilising mechanistic models of the data as a more reliable source of estimating behaviour in the extreme.

Acknowledgements

Before anyone else I need to acknowledge the unwavering support from my husband, Simon, over the years I have pursued this crazy dream. He has been there through all the high points and low, always cheering me on and believing in me. Without him, this would not have been possible. Thank you Simo, for being my rock, my net and my heart.

To Barbara Holland and Raimondo Bruno, I extend all my gratitude and thanks for your stellar supervision, for taking me and my project on even though the topics are on the fringes of both of your areas. Thank you for taking the risk and working with me, through everything the years have thrown at us, to get us to this point. You are two incredible role models and I hope that I can live up to the examples of collaboration, research and teaching excellence you have set.

Jeremy Pritchard, Simon Wotherspoon and Andrew Heathcote, your grounded and practical advice and support came at exactly the right time, more than once. Thank you all for taking the time to encourage and support me through my projects. And to John McKinlay, Steve Candy and the other contributors to Gentle R at the Australian Antarctic Division, thank you for being the statistical sounding board I needed and for providing inspiration and encouragement.

To the wastewater team at UQ, Foon Yin Lai, Phong Thai, Jochen Meuller and Jake O'Brien, thank you for trusting me with your data and welcoming me into your group. And to Christoph Ort I thank you for your wisdom and advice. I hope that we can achieve even more wonderful things together in the future.

For the people who have been there since the start, Jason Cosgrove and David Horsely, I appreciated you both so much. And to my office mates Minnie Mao, Steve Walters and Jonathon Mitchell thank you for your

feedback, friendship and encouragement. To the wonderful Karen Bradford, thank you for looking out for me and making sure I always had everything I needed so that I could keep on going throughout my candidature. Special thanks to all of the team in the School of Mathematics and Physics for your feedback and encouragement over the years and especially to Rhys Paul for your brilliant suggestion for the title of the thesis.

Most recently, thank you to my beautiful baby boy Benjamin. You have opened my eyes to a whole new world of amazing things and have provided the determination I needed to finish. Thank you also to my Mum, Coralie, and again to Simo for being there to help look after our little man Ben when I needed to focus and study. The support and love you have provided over these last few months have made more of a difference than you will ever know.

Finally, to my God-parents, Jean and Grant Brightman. Thank you for the hugs and the hard truths, and for believing in me. Without you I wouldn't have started.

Contents

Chapter 1

Introduction

This thesis explores how data is used to make inferences about different aspects of illicit drug use in society. Starting with an analysis of how measuring the extent illicit drug use can be best undertaken, and concluding with how the impacts of illicit drug use on an individual's cognitive processes can be statistically ascertained, this thesis considers a range of statistical techniques designed for use with temporally autocorrelated data. Throughout the process of applying these techniques, the theme of empirically driven analysis versus the use of mechanistic models to explain patterns in data is explored.

1.1 Empirical Analysis and Mechanistic Models

When observing complex data there are two general situations researchers can find themselves in. In the first situation, the researcher is observing a complex system of data in which little is known about the structure or functional mechanisms that have produced the data being observed. There may be hypotheses about some component of the underlying processes giving rise to the data, but these are often not enough to fully explain the variability in the real data. In this case analysis of the data is undertaken empirically, using regular, statistical, mathematical forms of analysis that are chosen based on the observed behaviour of the data (Thakur, 1991). Empirical analysis may take the form of a simple comparison of means using a t-test or sim-

ple models using linear regression; or it may be far more complex, using statistical models such as generalised additive models and time-series techniques using statistical methods such as maximum likelihood and Bayesian analysis techniques. For example, Peacock et al. (2013) completed repeated measures ANOVA to ascertain whether alcohol or energy drinks effect risk taking behaviour while Ort et al. (2014a) used time series analysis to explore changes in cocaine loads over time. Alternately, generalised additive models have been used as alternatives to time series for identifying patterns in data (Sullivan et al., 2015). The key aspect of all of these analyses, however, are that they *driven by the data* making no or few assumptions about how the data arose.

In the second situation, a researcher may have a well developed understanding of the system which created the observed data. More explicitly, the researcher may be able to define a model, based on their understanding of the mechanisms that produced the observed data, that can explain the main sources of variability in what has been observed. Thakur (1991) defines such a model as a

> "... a secondary system used to verify any hypotheses on the primary system." (pp. 41)

More explicitly, we consider that the observed data has been produced from a *primary system* which can explain all of the variability in the data. The researcher will not know all of the aspects of the primary system but can define some *secondary system* which explains the majority of the observed variability. This secondary system is the model which encapsulates the researchers understanding of the mechanisms producing the data. As such, the secondary system is a *mechanistic* model in that it is developed independently of the data and are based, instead, of the underlying mechanisms which produced it.

In the Philosophy of Science literature, the distinction between mechanistic and empirical models is focused around whether you are trying to use your model to explain your observations or predict future ones (Shmueli, 2010; Lehmann, 1990). Mechanistic models seek to explain the underlying processes that gave rise to the observations being studied, where as empirical models don't necessarily mind what created the observed data, just that they can predict how future data will look. There are, of course, benefits and limitations associated with both empirical analysis and the use of mechanistic models. Using well chosen empirical analysis, for example, will ideally find any trends and patterns within the existing data when the assumptions of the analysis are met. Mechanistic models are, in comparison, not necessarily going to represent the patterns in the data and, when this happens, can struggle to provide useful results when used. If the goal is to make inference about or explain the mechanisms underlying the observed data, a well-fitting mechanistic model will be able to achieve that goal. However empirical analyses, being driven by the data, do not necessarily have the same ability to link to the mechanisms underlying the observed data as a mechanistic model. The choice then of whether to use empirical analyses or a mechanistic model appears to be driven by two main features:

1. Whether the analysis aims to make inferences about the underlying mechanisms that created the data or the measured data itself.

2. Whether there is enough knowledge of the underlying mechanism to create an effective mechanistic model.

So, despite the apparently clear distinction between the empirical and mechanistic, practically the line between these two techniques is not so well defined. When the second point has not been met, empirical analysis can be used to inform about sources of variability not yet captured by mechanistic

models driving future scientific development (Shmueli, 2010). As such, we see empirical analyses informing the creation of mechanistic models. In the field of psychological science, Ferguson (2015) appears to agree with the use of empirical methods to drive mechanistic models. They argue that theories (such as those about different aspects of cognitive processing and behaviour) can become so well established that they can be seen more like a "truth" than the heuristics that they are. Finding this new evidence to update these models is challenging, however, if the only means of analysing data is through the use of a popular mechanistic model. Ferguson (2015) suggests then, that to find new evidence, theoretical models which are less mechanistic are required in the field of psychology.

In a direct reply to Ferguson (2015), Tryon (2016) argues against a move toward empirical analyses and instead argues for more mechanistic models, richer in nature, to compete with the existing models in the literature. And it is perhaps here that we see the cross over between the practical understanding between mechanistic and empirical analysis. Lehmann (1990) points out that, even when using empirical analysis such as regression, choosing an appropriate ARMA model or deciding on the number and form of interactions in a factorial model the researcher must have first chosen the class of models they wished to apply. Although, in some cases, this choice will be very clear given the observed data, in others the choice may be driven by more mechanistic considerations. So when does this process become more mechanistic than empirical? Despite the differences in the suggestions by Ferguson (2015) or Tryon (2016), the common theme of remaining flexible and maintaining a search for better ways of explaining and describing observed psychological phenomena remains constant, and this thesis has the same general aim. The following chapters present both empirical and mechanistic analyses of a range of psychological data with the goal of improving

the information we can glean and the broadening the applicability of any findings.

1.2 Temporal Autocorrelation

A common theme of all of the data considered in this thesis is temporal autocorrelation. Data which has a temporally autocorrelated structure violates the underlying assumption of many common statistical techniques. Regression, for example, assumes that each point is independent; which is clearly violated in temporally autocorrelated data in which observations that are closer together are more "alike" than those farther apart. Because of this lack of independence, analysis techniques that take into account this serial dependence need to be employed.

There are several ways of dealing with temporal autocorrelation which are driven by the type of data collected and the research questions being asked. For example, longitudinal data, collected at regular intervals, spanning a long range of time is often analysed using time series analysis or generalised additive models while repeated measures data over shorter periods can be analysed using one-step ahead predictions of individual performance. In addition to the styles of analysis chosen, the fundamental theory of probability that a statistician adheres to can also differ. Bayesian and frequentist techniques differ in their fundamental treatment of probabilities and that leads to very different styles of analysis. Silva (2017) explains that Bayesians are concerned with correctly interpreting information that has been updated post-experiment while frequentists are focused, before the experiment begins, on controlling the probability of coming to the wrong conclusion. This difference, although seemingly semantic in its simplest form, has real implications for analysis and can often lead to vastly different conclusions being drawn from the same data. Increasingly researchers will

analyse their data under both frameworks (a simple Google Scholar search on the 30^{th} of September 2017 of the term "frequentist and Bayesian" revealed 807 results for 2017 alone while a search for "frequentist" and "Bayesian" returned 4,060 results) and the debate over which techniques are most accurate is still a vibrant one.

This thesis explores a range of possible analytical techniques that cover Bayeisan and frequentist frameworks in both mechanistic and empirical analysis styles. This is both in an attempt to showcase some of the strengths and weaknesses of the multiple forms of analysis, and to demonstrate the breadth of analysis techniques available for dealing with temporally autocorrelated data.

1.2.1 Time-Series

Time-series analysis is a common way of analysing data that has been collected sequentially at equally spaced time intervals over a long period of time. Importantly, this *long period* of time is relative to the intervals over which the data were collected. For example, data collected annually will require many years of observations, whereas data collected every minute may only need a few days of data to be useful (Cowpertwait and Metcalfe, 2009).

In its simplest form, time series uses the features of the data to iteratively inform the model of the data, making this an empirical method of analysis. The main features of a time series are usually an overall *trend* and, potentially, some sort of *seasonal* variability and it is the aim of analysis to extract and/or measure these two features. The trend and seasonality then provide a more concise summary of the characteristics of the time series that can be used for understanding how the measured observations have changed in the past and, potentially, forecasting how they may continue to change into the future (Cowpertwait and Metcalfe, 2009). It is therefore

the features of the observed data, rather than the mechanisms *creating* the observed data, that are of most interest when using time series analysis. As such, the statistical model itself is not of great interest as the trend and seasonality components (along with the analysis of any remainder) contain all the information relevant to the research question.

If the research question is focused more on the mechanisms creating the data, rather than the data itself, then time-series analysis may no longer be an appropriate method of analysis. Any model used to fit the data ideally would have parameters that can be linked to or interpreted in light of the underlying mechanisms which gave rise to the data. Predicting decision-making in the face of uncertainty, for example, may be predicted using the outcomes of previous decisions but it may also be the case that the interest lies in understanding the cognitive processes that led to the decisions being made rather than the outcome itself. In this case, incorporating information about cognitive processing into a model of the observed behaviour would be essential to be able to estimate how the cognitive processes are behaving or changing over time. Time series analysis does not provide such models. In these cases, mechanistic models that are constructed using knowledge about the processes underlying the observations are favoured.

1.2.2 One-step Ahead Prediction

Mechanistic models designed to explain or predict temporally autocorrelated data must take into account the lack of independence between sequential observations in the data. This can be achieved by using one-step-ahead prediction where the model is developed to predict the next observation in a time series based on the sequence of measured events up to and including the current observation. Examples of such models are the Expectancy Valence Model (presented in Chapter 4) and the two-parameter BART model

(presented in Chapter 6).

When considering a mechanistic model, the goal is to estimate the parameter values of the given model for either a group or for an individual data set with the aim of making inference about the underlying mechanisms which created the data. The techniques used to find the optimal parameter estimates given the data are often maximum likelihood or Bayesian in nature and both of these styles of analyses will be explored in this thesis.

1.3 Drug use and its Impacts on Society

The final common thread weaving through the chapters of this thesis is that of illicit drug consumption. The use of illicit drugs is an area of interest across a broad range of industries and fields including public policy, law enforcement and physical and mental health. Accurately measuring the extent of illicit drug consumption in the community, along with an understanding of its impacts on the health and well-being of public, are fundamentally essential to providing the services required to both regulate use and provide treatment options for rehabilitation. All of the data sets considered in this thesis are related to illicit drug consumption with the goal of analysis to better understand either the prevalence in the community or impact of illicit drug use on an individual's cognition.

Both the field of measuring drug consumption in the community and understanding its cognitive impacts are therefore the subject of constant research, development and innovation. One such recent innovation has been in the area of measuring illicit drug consumption extent through the sampling of drug residues in sewerage (wastewater). Although the extent of illicit drug consumption across a population has historically been ascertained using techniques such as surveys, crime statistics, drug seizures and consumer interviews (Castiglioni et al., 2006; Banta-Green et al., 2009; Chen

et al., 2011), limited population coverage (limited access to populations such as those who are incarcerated or without telephones) and self-report bias impact on the validity and reliability of this collected data (Banta-Green et al., 2009). The Illicit Drug Reporting System and the Ecstasy and Related Drugs Reporting System are two surveys routinely used in Australia to monitor drug use which provide critical information on illicit drug consumption extent (Chen et al., 2011), but the fast moving trends associated with illicit drug consumption are also hard to monitor in real time using survey type methods due to the amount of time required to collect and process information (Banta-Green et al., 2009).

Using a technique developed by Zuccato et al. (2005), analysing the levels of illicit drug metabolites from a wastewater treatment plant attempts to solve these problems. Analysis of wastewater is conducted using high-pressure liquid chromatography tandem mass spectrometry (Castiglioni et al., 2006) which can provide relatively timely results and, as a wastewater treatment plant is considered a closed water system in that it serves a known population (Frost et al., 2008), those results provide a snapshot of an entire *region* rather than a specific demographic. For this reason, once allowances have been made for the pharmacokinetics associated with a specific drug, levels of the drug or its major metabolites can be seen as an indicator of the level of consumption in the contributing population.

1.3.1 Waste-Water Analysis and Empirical Models

Wastewater-based epidemiology (WBE) is the process of extracting metabolic residues from wastewater in order to make inferences about the characteristics of the contributing population. It involves taking regular samples (usually liquid) from a wastewater treatment plant and, after some form of storage by freezing, they are analysed using mass-spectrometry to measure

compounds of interest within the sample. For example, when using illicit drugs, the drugs are metabolised and any remaining metabolised version (metabolite) of the drug is excreted. The amount of metabolites in the wastewater can then be measured to give an estimate of the amount of parent compound that may have been consumed to create it. This emerging technique of using waste-water analysis to measure illicit drug consumption in communities was first utilised in 2005 (Zuccato et al., 2005) and has undergone rapid acceptance as a way of gaining increasingly accurate information all around the world (for examples see Zuccato et al. (2016); Bijlsma et al. (2016); van Wel et al. (2016); Been et al. (2016); Lai et al. (2016b)). This is, however, a relatively new technique with rapidly changing markers of best practice (Ort et al., 2014a; Thai et al., 2016; Humphries et al., 2016; Ort et al., 2014b; Zuccato et al., 2016). With greater understanding of the samples themselves is coming a greater understanding of how much information can actually be gleaned from samples.

Due to the resource intensive nature of the collection and analysis process, many early studies have relied on few samples taken over short periods of time (van Nuijs et al., 2009; Prichard et al., 2012; Zuccato et al., 2011; Khan et al., 2014; Kim et al., 2015) but as the popularity of this technique has increased, so too has the number of studies looking to take longer term, more intensive samples (for example van Nuijs et al. (2011b); Ort et al. (2014a); Lai et al. (2016a)). Chapter 2 explores how best to undertake longitudinal monitoring so as to maximise the information obtained whilst minimising the associated costs.

1.4 Thesis Structure

1.4.1 Chapter 2

Chapter 2 presents an edited version of a published paper which combines the original article and it's supplementary material into a coherent whole (Humphries et al., 2016). Based on longitudinal illicit drug consumption data, collected from a waste-water treatment plant over 13 months between May 2011 and June 2012, the data were analysed to estimate overall trend in illicit drug consumption as well as estimates of monthly levels of use. Conducting longitudinal monitoring programs to investigate temporal trends in drug consumption is challenging as it is very costly and resource intensive. Prior to this publication, the only work in this area was a study by Ort et al. (2014a) which investigated how a stratified random monitoring scheme can allow reliable estimation of the average yearly usage of cocaine in a small catchment. There remained a significant research gap to examine how to design monitoring schemes to provide a sound assessment of temporal changes in use for different types of drugs. Chapter 2 fills that gap.

In Chapter 2 time series analysis is used to evaluate the accuracy and robustness of a range of potential monitoring schemes in order to select a cost-effective scheme that still allows reliable interpretations of temporal trends of illicit drug usage. For the first time such a study was carried out using data ($\sim$ 350 days) obtained in a large catchment and for four different drugs. It was found that:

- It is important to understand the weekly cycle of drug consumption before designing the sampling scheme.
- Utilising the weekly cycle information to inform a monitoring strategy maintains the ability to reliably interpret temporal trends while minimising the cost of data collection.

The findings reported in Chapter 2 are important because, for the first time, it provides researchers with a comprehensive assessment of monitoring schemes for drugs with different usage patterns. Such information will help researchers in the field to design long term monitoring of drug consumption in different sewage catchments in a cost-effective way.

The comprehensive assessment of strategical monitoring schemes presented was an interdisciplinary achievement among statisticians, sewer engineers and environmental scientists. The data was provided by the environmental scientists Foon Lai, Phong Thai, Jake O'Brien and Jochen Mueller who both collected the samples and completed the chemical decomposition (as presented in their article Lai et al. (2015)). All statistical analyses were completed by me with guidance and suggestions from Barbara Holland and Christoph Ort. I also wrote the majority of the article with guidance from Raimondo Bruno and editing feedback from all authors. Foon Lai and Phong Thai provided the background information for Section 2.2.1 in the Experimental Methods.

Chapter 2 focuses on empirical analysis of time-series data. Time series analysis is a frequentist style of analysis that provides visual estimates of observed trends and empirical models makes sense in this case, where the goal is to understand and make predictions about the exact thing that is being measured. Although time series analysis is widely used for longitudinal data, there is some evidence to suggest the use of generalised additive models can provide results that allow stronger inference than the general trends reported in time series. This is discussed as a direction for future work in Chapter 2. Although there are Bayesian techniques for time series analysis (e.g. see West et al. (1985)) they are not yet widely used and so are not considered here.

In the field of mathematical psychology, however, Bayesian techniques

are widely used. When looking at the impact of drug use on cognitive function, even though we may still have temporally autocorrelated data, more information than general trends in outcomes are required and so alternatives to time series analysis must be explored.

1.4.2 Chapter 3

Chapters 4-6 consider the impact of drug use on cognitive function through the use of both empirical and mechanistic models to temporally autocorrelated data. Chapter 3 provides an introduction to cognitive models and the utility of empirical and mechanistic styles of modelling in this space as well as a breif discussion of the features and benefits of Bayesian and frequentist analysis styles.

1.4.3 Chapters 4 and 5

Chapter 4 presents a published paper (Humphries et al. (2015) with supplementary material in Chapter 5) and addresses whether the Expectancy Valence Model (EVM), a mechanistic model of cognitive performance, is suitable for use at decomposing behaviour on the Iowa Gambling Task (IGT), a psychological assessment tool, at the level of the individual. The data analysed in Chapter 4 was provided by Raimondo Bruno and collected by Jessica Hartley and Elizabeth Murray. The analysis presented in Chapter 4 was completed by me with guidance from Simon Wotherspoon. The Chapter itself was written in it's entirety by me with guidance from Raimondo Bruno and Simon Wotherspoon with editing feedback from Yuliya Karpievitch.

The EVM of the IGT is a commonly used mechanistic model aimed at identifying the underlying psychological processes responsible for decision making deficits and can be purchased for use in clinical practice (Bechara, 2012; Buelow and Suhr, 2009). However, high levels of uncertainty surrounding estimates of cognitive performance at the level of the individual have

been reported by a growing body of literature (Wetzels et al., 2010; Steingroever et al., 2013a,b). Chapter 4 confirms that the EVM does not provide clear information about decision making processes at the individual level by fitting the EVM, with individual random effects, to a sample of participants from various drug using populations using Bayesian techniques. The uncertainty associated with parameter estimates, gained using the EVM, are revealed to be due to non-normal and unreasonably localised estimates with observed estimates showing bi-modality, non-linearity or spanning the entire parameter space. This not only makes analysis more difficult but renders psychological interpretation of these estimates at the level of the individual contradictory or even misleading. Using these sorts of estimates to inform clinical practice at the level of the individual can therefore lead to treatment recommendations that are not appropriate for the individual being measured.

A novel second experiment, where participants are asked to complete the IGT multiple times in an effort to increase sample size, also shows an inability to gain precise estimates of cognitive performance for the same non-normal and unreasonably localised reasons. But, in an attempt to increase the validity of individual-level parameter estimates, a third experiment proposing a new, two-parameter version of the EVM is presented. In the two-parameter implementation of the EVM, results were clearer and more easily interpretable than when using the traditional EVM suggesting this may be more viable for use in a clinical setting.

1.4.4 Chapter 6

Chapter 4 highlighted that the IGT requires a complex system of cognitive processes to produce behaviour and that the parameters of the EVM were not able to model these cognitive processes accurately at the level of the

individual. The solution presented in Chapter 4 was to simplify the model, but Chapter 6 presents another solution: Using a cognitive assessment tool which relies on a less complex system of cognitive processes to produce behaviour.

Chapter 6 presents simulation studies of the Balloon Analogue Risk Task (BART) which is a simple cognitive assessment tool designed to assess risk seeking behaviour alone (Lejuez et al., 2002). Particular importance is paid to whether the behaviour of individuals who are expected to have cognitive deficits of some kind can be well recovered by the mechanistic and empirical models considered. This focus links clearly with the clinical applicability of the BART and its associated models as individuals presenting for clinical assessment are likely to have cognitive deficits that give rise to behaviour outside of, or more extreme than, what is considered normal.

Three models of performance on the BART are presented, the mechanistically derived van Ravenzwaaij et al. (2011) two-parameter BART model, and two new empirical models, The Basic Response Model (BRM) and the Run Dependent Response Model (RDRM). Data was simulated and parameter recovery, using maximum likelihood, completed for each model respectively to ascertain if parameter estimates of any model of behaviour on the BART could be gained with enough accuracy to be useful at the level of the individual.

Chapter 6 suggests that the van Ravenzwaaij et al. (2011) two-parameter BART model may not always provide reliable results if used in a clinical context where the individual being measured has some cognitive deficit. The RDRM was able to recover parameters well in most cases and the BRM gave the most reliable estimates of all the models considered. Given that the BRM and RDRM are both empirical models, however, links between the model parameters and the cognitive processes they may represent need

to be established before their use in a clinical context. Chapter 6 therefore suggests validity studies and application of the BRM and RDRM to real data as future endeavours.

Chapter 2

Evaluation of monitoring schemes for wastewater-based epidemiology to identify drug use trends using cocaine, methamphetamine, MDMA and methadone

This Chapter presents a version of the published article:

Humphries, M. A., Bruno, R., Lai, F. Y., Thai, P. K., Holland, B. R., O'Brien, J. W., Ort, C., and Mueller, J. F. (2016). Evaluation of monitoring schemes for wastewater-based epidemiology to identify drug use trends using cocaine, methamphetamine, MDMA and methadone. *Environmental Science & Technology*, 50(9):4760 - 4768.

The following chapter presents the article with supplementary material integrated into the body of the work, along with some extensions.

Chapter 3

Modelling Cognitive Performance

The purpose of Chapter 2 was to look at ways of increasing the efficacy of measuring illicit drug use using wastewater data. However, the extent of illicit drug use in the community is not the only area in which temporally autocorrelated drug use data can be found. The following Chapters will consider temporally autocorrelated data aimed at measuring the effect of illicit drug use on an individual's cognitive performance.

3.1 Modelling Cognitive Processes

Most commonly, to measure decreases in an individual's cognitive performance or *cognitive deficits*, an individual is given some cognitive assessment tool (in the form of a questionnaire, test or game) to complete. The resulting answers or behaviours displayed by the individual are driven by the cognitive processes producing that behaviour so by measuring the behaviour, indirect measurements of the cognitive processes are obtained.

In the case of a game, an individual may be asked to complete a task repeatedly so as to gain a large volume of data from which to estimate performance. Chapters 4 and 6 consider two such games; The Iowa Gambling Task (IGT; Chapter 4) and the Balloon Analogue Risk Task (BART; Chapter 6). In the IGT, individuals are asked to repeatedly choose from decks of cards to win money and their choices are directly related to how much

money they win. The BART has a similar goal of winning as much money as possible, but to do so, the individual is asked to repeatedly pump up balloons as large as possible without bursting them (larger balloons make more money).

The IGT and the BART both use a gambling-style, money maximising scenario, and so are commonly used for ascertaining whether individuals are highly impulsive. High levels of impulsivity are theorised to be related to continual use of some illicit drugs (Belin et al., 2008; Baldacchino et al., 2015; Ansell et al., 2015) making these assessment tools particularly relevant for measuring the effects of drug use at the level of the individual. Research into the relationship between illicit drug use and impulsivity, as both a predictor and a consequence of continued illicit drug use, is ongoing (Moeller et al., 2001; Dawe and Loxton, 2004; de Wit, 2009; Nuijten et al., 2016; Kluwe-Schiavon et al., 2016). But evidence has shown that highly impulsive individuals are more likely to relapse, when trying to abstain from illicit drug use, than their less-impulsive counterparts (Moeller et al., 2001; Economidou et al., 2009; Nuijten et al., 2016). In a clinical setting understanding whether a client is highly impulsive and therefore less likely to present for treatment (Moeller et al., 2001) or whether they are more likely to discount risk over reward (Dawe and Loxton, 2004) can have implications for type of treatment suggested. Incorporating techniques for dealing with high risk situations and encouraging attendance at treatment will therefore be an integral part of maximising recovery for highly impulsive individuals. Identifying impulsivity at the level of the individual, with a good level of accuracy, is therefore the focus of Chapters 4 and 6.

3.2 Empirical and Mechanistic Models of Cognitive Performance

Whether considering the IGT or the BART, participants are asked to complete many sequential trials on the game. Completing many trials allows for averaging of performance across a greater number of trials, getting closer to what might be seen as *true* performance estimates. However, when completing multiple trials on the same task, there is clearly sequential information being processed by the individual leading to a lack of independence between trials. For both games then, the temporal autocorrelation in the data from each individual is a factor that must be considered in analysis. However, unlike in Chapter 2, time series analysis may not be the best way to model this sort of data.

As described in Chapter 1, time series analyses breaks down autocorrelated data into an overall trend and, potentially, some sort of seasonal variability. Given that the aim of any analysis of the IGT and BART data is to gain some measure of cognitive performance, with a focus on when performance on the task changes rapidly in response to negative or positive outcomes, trends and seasonal effects produced by time series analysis does not match well with answering this aim. Empirical analysis then usually focuses on summaries of the data such as average amounts of money won, average number of times completing options are taken (for example Kim et al. (2016); Dai et al. (2015)) or, in the case of the BART, average number of time balloons are pumped up (for example Kessler et al. (2016); Balaguero et al. (2016)). These sorts of analysis are capable of providing summaries of performance but, just like times series, they do not directly model the underlying cognitive processes that gave rise to the observed behaviour.

To make inferences about the cognitive processes that give rise to the measured behaviour, mechanistic models driven by the theories about how

the underlying cognitive processes work are used. Chapters 4 and 6 consider two popular mechanistic models which are implemented in the IGT and the BART respectively. Chapter 4 (and the associated supplementary material in Chapter 5), presents a published article in which a mechanistic model is suggested as an alternative to a commonly used mechanistic model for decomposing behaviour on the IGT (Humphries et al., 2015). The analysis explored covers some empirical summaries but uses Bayesian analysis techniques to ascertain the ability to gain insight into cognitive deficits at the level of the individual.

Extending on this, Chapter 6 also looks at cognitive deficits at the level of an individual, but presents a simulation study comparing empirical model and mechanistic models used to describe performance on the BART. In Chapter 6, we contrast a leading mechanistic model with a purely empirical model using frequentist model comparison techniques. We also propose a mechanistic extension of the empirical model, based on our understanding of behaviour rather than driven by the data, which illustrates one of the possible intersections between these two styles of analyses.

3.3 Frequesntist and Bayesian Analysis

The differences between Chapters 4 and 6 do not stop at the level of empirical versus mechanistic modelling. They also deal with two very distinct ways of analysing data covering the application of Bayesian and frequentist techniques. In the field of mathematical psychology, Bayesian styles of analysis in cognitive modelling is extremely common due to both the complexity of the models and the cognitive processes they represent.

3.3.1 Maximum Likelihood

Both frequentist and Bayesian analyses use maximum likelihood techniques, in some form, to reach their conclusions but it is in how maximum likelihood

is used specifically that these two theories differ. Maximum likelihood is a classical, frequentist, statistical estimation technique in which inference is based on the likelihood of the data alone (Congdon, 2003). Specifically, Maximum likelihood estimation provides the most likely point estimate for a parameter of interest given the observed data (Pawitan, 2001).

Let $Y_1, ..., Y_n$ be independent samples drawn at random for a distribution with probability density function $f(y|\theta)$. Here the y is a vector of observed data and θ is the vector of parameters in the distribution function of y. If we then view f as a function of θ that is conditional on the observations y, we obtain the likelihood function

$$L(\theta|y) = \prod_{i=1}^{n} f(y_i|\theta). \tag{3.1}$$

Maximum likelihood estimation is the process of determining the value of θ that maximises the likelihood function.

The Maximum Likelihood Estimator (MLE) is the $\hat{\theta}$ that maximises the log likelihood is obtained by solving when the score function

$$S(\theta) \equiv \frac{\partial}{\partial\theta} \log L(\theta) = 0.$$

To ascertain the variance associated with the MLE, the *observed Fisher information*

$$I(\theta) \equiv -\frac{\partial^2}{\partial\theta^2} \log L(\theta)$$

must be considered. Large curvature in $I(\hat{\theta})$ is associated with a tight peak, indicating more knowledge about θ. The law of large numbers implies that as $n \to \infty$, the maximum likelihood estimates will be approximately normally distributed with a standard error determined by the inverse of the Fisher Information. This can be seen by expanding $\log L$ in a Taylor Series about the MLE

$$\log L(\theta) \approx \log L(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^{(T)} I(\hat{\theta})(\theta - \hat{\theta}). \tag{3.2}$$

The first order term of the expansion is zero at a maximum and, for large samples, neglecting the larger order terms is justified by the law of large numbers so that approximately

$$\theta \sim N(\hat{\theta}, I^{-\frac{1}{2}}(\hat{\theta})).$$

So parameter estimates are considered to be approximately normally distributed and the Fisher information gives the inverse covariance of parameter estimates. This argument relies both on a quadratic approximation and large samples, but the validity of the Taylor Series approximation also needs to be good in the vicinity of the mode (i.e. a *single* mode). This means that if the MLE is near the boundary of parameter space, or if there are multiple modes, the arguments fall apart. Simple point estimates and confidence intervals in these cases are not meaningful when obtained in this way.

For any finite sample, there is no guarantee that normality assumptions are accurate for non-linear problems, such as with many cognitive models. Further to this, estimates using the MLE can be inaccurate for parameters with bounded domains (Pawitan, 2001), as the Taylor Series argument in Equation 3.2 is called into question and, in mechanistic, cognitive models it usual to see parameters with bounded. So it would appear that, for mechanistic cognitive models, maximum likelihood estimation can run into a raft of problems that can call it's results into question. Bayesian estimation techniques do not carry these assumptions and can quite easily overcome problems associated with non-linear modelling and bounded parameter domains which is why it is so widely used in the field of mathematical psychology.

3.3.2 Bayesian Analysis

Unlike maximum likelihood techniques, which rely on the long run frequency of events to define probability, the Bayesian framework is based on degrees of belief. This means that, instead of estimating fixed model parameters once

the data has been observed, Bayesian estimation allows us to consider our parameters as random variables and build up our knowledge about how they are distributed as we see evidence. In Bayesian estimation we are estimating the posterior density $P(\theta|y)$ of our parameters θ

$$P\left(\theta|y\right) = \frac{P\left(y|\theta\right)P\left(\theta\right)}{\int_{\theta} P\left(y|\theta\right)P\left(\theta\right)} \tag{3.3}$$

where $P(y|\theta)$ is the likelihood of our observations y given θ and $P(\theta)$ is our prior distribution. The prior distribution is a unique feature of Bayesian analysis and encapsulates our knowledge about how our parameters are distributed. This is usually chosen to be uninformative and it is updated once the evidence has been observed to approach the true distribution of θ. In this way, Bayesian techniques avoid the problems with Newton minimisation described above by providing the opportunity for deriving the true distribution of parameters without forcing them to adopt a normal distribution.

The flexibility of Bayesian analysis to allow model parameters to define their own distributions is particularly important with mechanistic models that have complex relationships between parameters. This is the main appeal of these methods in the field of mathematical psychology. However, these techniques have not been quite so emphatically adopted in other areas where models tend to meet the assumptions of traditional, frequentist forms of analysis. This is particularly true of empirical, data driven models which often have less complex relationships between parameters. This thesis looks at an example of when Bayesian analysis is particularly necessary in the analysis of the Expectancy Valence Model of the Iowa Gambling Task in Chapter 4. In this chapter we present a published paper showing how a complex, mechanistic model of cognitive performance fails the assumptions required for maximum likelihood estimation through a full hierarchical Bayesian analysis of the model.

In Chapter 6 another mechanistic model of cognitive performance, which measures outcomes on the Balloon Analogue Risk Task, is presented. Again, it is shown that the assumptions for maximum likelihood estimation are not met, but this time the conclusion is reached through a simple frequentist analysis of the model. In addition, two simple empirical models of performance are suggested and are analysed using frequentist methods with success. This perhaps highlights the reasoning behind the lack of rigorous uptake of Bayesian methods across fields who rely heavily on empirical techniques.

Chapter 4

The Expectancy Valence Model of the Iowa Gambling Task: Can it produce reliable estimates for individuals?

This Chapter presents an edited version of the published article:

> Humphries, M. A., Bruno, R., Karpievitch, Y., and Wotherspoon, S. (2015). The expectancy valence model of the Iowa gambling task: Can it produce reliable estimates for individuals? *Journal of Mathematical Psychology*, 64 - 65:17 - 34.

4.1 Introduction

Impulsivity is a multidimensional construct; it can, for example, relate to a failure to follow instructions or wait for one's turn (execution impulsivity); responding before all the essential information has been gathered (preparation impulsivity), or failing to delay gratification; focussing on short term or positive outcomes and relatively discounting long term or negative outcomes (outcome impulsivity) (see Evenden (1999b) for a review). Developed by Bechara et al. (1994), the Iowa Gambling Task (IGT) is a four-armed bandit task designed to measure deficits in decision making among clinical populations, in particular the notion of outcome impulsivity. To complete the IGT, a participant chooses from four computerised decks of cards to try

and maximise their long-term return. Successful completion of the task requires the participant to learn that two of the decks are disadvantageous over time (high immediate returns but long term losses) while the remaining two decks are advantageous (low immediate win amounts but long term gains). Highly impulsive individuals will, theoretically, show poor performance on the IGT due to the appeal of the high immediate win amounts associated with the disadvantageous decks (Bechara et al., 1994).

The IGT is currently being sold as a clinical assessment tool for the assessment of individual decision making deficits (Bechara, 2012; Buelow and Suhr, 2009). With multiple, independent studies showing an association between IGT scores and substance use relapse (Nejtek et al., 2013; De Wilde et al., 2013b; Radat et al., 2013; De Wilde et al., 2013a; Wang et al., 2013; Goudriaan et al., 2011; Kasar et al., 2010; Salgado et al., 2009; Passetti et al., 2008), the IGT is a popular choice when developing treatment plans in a clinical setting. Two of the simplest, standard ways of identifying poor performance on the IGT are to examine the overall net return after a specified number of trials, or to look at the frequency of choices from advantageous and disadvantageous decks in blocks across the duration of the task (Poletti et al., 2011; Brown et al., 2012; Lamers et al., 2006; Stout et al., 2004). However, some studies have demonstrated that these standard measures have questionable validity (Lin et al., 2013; Steingroever et al., 2013a; Buelow and Suhr, 2009). More pertinently, both net return and frequency of deck choice measure composites of multiple decision making processes, making it hard to argue that poor performance indicated by these measures is due to impulsive behaviour alone. For example, a participant has to remember multiple outcomes over time in order to make the most advantageous choices in the future, so it is reasonable to think that deficits in learning processes will also lead to poor performance on the IGT due to

that fact.

To disentangle the differences between poor performance on the IGT due to high outcome impulsivity or due to poor learning, a more sophisticated mode of analysis is required. Cognitive models of performance allow the underlying psychological processes driving observed performance to be teased apart and measured. In this way behaviours that are composites of different psychological processes can be understood in greater depth. Although several cognitive models have been proposed to disentangle the psychological processes underlying performance on the IGT (for examples see Steingroever et al. (2013b) and Ahn et al. (2008)), the Expectancy Valence Model (EVM) proposed by Busemeyer and Stout (2002), has been the most widely implemented. Producing estimates of impulsivity or motivational processes, memory and learning, and response consistency (Busemeyer and Stout, 2002), the EVM has been successfully used to identify high impulsivity levels in cocaine users (Stout et al., 2004), memory deficits in Huntington's sufferers (Busemeyer and Stout, 2002) and differences in a raft of other psychological groups of interest when compared to control groups (for a review of the applications of the EVM on the IGT see Yechiam et al. (2005)).

Given the success of the EVM at decomposing performance on the IGT for groups, and given that the IGT is widely used in clinical settings, there would be great potential benefits in applying the EVM of the IGT at the level of the individual. For example, individuals with drug addictions are known to be more likely to relapse following rehabilitation if they are highly impulsive (Nejtek et al., 2013; De Wilde et al., 2013b; Radat et al., 2013; Passetti et al., 2008). Using the IGT as an assessment tool, and then decomposing behaviour for the individual using the EVM, a clinician would get a clearer estimate of impulsivity for the individual than using the standard, composite measures of performance on the IGT such as net return. If a client

was identified as being highly impulsive, then any treatment plan could include extra coping strategies for high risk situations to try and avoid relapse. However, this depends on gaining a *clear and valid* estimate of impulsivity for the individual.

An increasing body of literature shows that estimates using the EVM to identify deficits in the psychological processes required to complete the IGT at the level of the individual produce highly uncertain estimates. Wetzels et al. (2010) highlight this problem by showing that EVM parameter estimates are highly uncertain at the level of the individual, even for simulated data in which the parameter values are known. Without precise estimates, it would be inappropriate to use the EVM to decompose behaviour on the IGT at the level of the individual and there are warnings in the literature against this course of action (Wetzels et al., 2010).

If it were possible to reduce the uncertainty associated with EVM estimates of impulsivity and memory when the IGT is used as an individual assessment tool, the EVM could be used to identify potential cognitive deficits leading to poor performance at an individual level. These results would provide more clinically useful information than the composite measures currently in use and would assist in tailoring treatments specific to the needs of the particular person. However, to be able to reduce uncertainty, the reason for the existence of the uncertainty must be determined. In this paper, we aim to explore why the EVM produces such highly uncertain individual-level parameter estimates, and examine an option for reducing uncertainty.

4.1.1 The Iowa Gambling Task

Proposed as a simulation of real-life decision making in the face of uncertainty, the Iowa Gambling Task (IGT) requires participants to make a series

of choices from four virtual decks of cards with the aim of maximising the amount won. The four presented decks have fixed (but undisclosed) win-to-loss ratios and dollar amounts, with two decks culminating in overall wins and two in overall losses (Table 4.1). Participants with unimpaired decision making processes converge to choices from profitable decks only (Busemeyer and Stout, 2002). A full description of the task is available in Wetzels et al. (2010).

Table 4.1: Payoff scheme of the traditional IGT (Bechara et al., 1994). Decks A & B may yield higher reward amounts but their associated loss amounts are also larger, resulting in net losses if chosen regularly. Decks C and D are, therefore, considered the advantageous decks.

	Bad Decks		Good Decks	
	A	B	C	D
Reward/Trial	100	100	50	50
Number of Losses/10 cards	5	1	5	1
Loss/10 cards	-1250	-1250	-250	-250
Net outcome/10 cards	-250	-250	250	250

Successful completion of the IGT requires the participant to explore all of the decks and, once all of the decks have been thoroughly explored, exploit the most profitable decks. To achieve this goal, a participant must evaluate the outcome of every deck choice, use this information to update any expectancies about returns associated with the decks and then make subsequent decision based on what has been previously learned. It is proposed that distinct brain regions or systems are responsible for producing each of these three processes and, as such, performance levels in each one of these processes can be depleted or vary independently of performance the others (Stocco et al., 2009). Poor performance on the IGT may be interpreted as a possible deficit in the relevant process.

4.2 Theory

4.2.1 The Expectancy Valence Model

The EVM is a cognitive model designed to interpret performance on the IGT by identifying the underlying psychological processes responsible for deficits in decision making. This utilises a series of four nested equations that attempt to model a respondent's motivational processes, memory/learning and consistency of choice (Busemeyer and Stout, 2002). Several assumptions are used by the EVM. First, the perceived value of a return may be different from the actual net value (Busemeyer and Stout, 2002). A *Valence*, in the form of a weighted, unidimensional utility function, incorporates this assumption into the model when a deck has been chosen. The Valence v is defined as

$$v_{i,t} = W \times win_t + (1 - W) \times loss_t \tag{4.1}$$

for the chosen deck i on trial t, where $0 < W < 1$ is the weighting parameter, win_t is the amount won on trial t and $loss_t$ is the amount lost on trial t. The weighting parameter W therefore represents shifts in attention between wins and losses, theoretically differentiating between highly impulsive individuals, who focus heavily on wins, and more conservative individuals that have a greater focus on the implication of losses (Busemeyer and Stout, 2002).

Second, any expectations about the outcomes associated with choosing a deck are learned by experience (Busemeyer and Stout, 2002). Once the Valence has been formed, it is incorporated into the body of knowledge an individual is building about a deck forming the *Expectancy*

$$E_{i,t} = (1 - \phi) \times E_{i,t-1} + \phi \times v_{i,t} \tag{4.2}$$

for deck i on trial t, where $0 < \phi < 1$ is the weighting parameter which determines the updating rate of the Expectancy. It is important to highlight here that the Expectancy of deck i is only updated when deck i is chosen.

The Expectancy for the remaining three decks will remain exactly the same until they are chosen again at some point in the future.

Consider a sequence of trials T for a given deck i. Then T_n denotes the n^{th} time, irrespective of trial number, that deck i was chosen. For example, when deck i is chosen for the third time, $n = 3$ such that T_3 indicates the third time deck i was chosen. Equation 4.2 can therefore be written as

$$\begin{aligned} E_{i,T_n} &= \phi \times v_{i,T_n} + \phi(1-\phi) \times v_{i,T_{n-1}} + \phi(1-\phi)^2 \times v_{i,T_{n-2}} + \dots \\ &\quad \dots + \phi(1-\phi)^{T_{n-1}} \times v_{i,T_1} + (1-\phi)^{T_n} \times E_{i,0}. \end{aligned}$$

Here it is easier to see that for values of ϕ close to zero, the updated Expectancy will be dominated by the initial Expectancy $E_{i,0} = 0$, which is a state of no knowledge.

The series representation of the Expectancy can be summed to yield an explicit expression for the Expectancy

$$E_{i,T_{n+1}} = (1-\phi)^{T_{n+1}} \left(E_{i,0} + \sum_{k=0}^{n} \frac{\phi \times v_{i,T_k}}{(1-\phi)^{T_{k+1}}} \right) \tag{4.3}$$

for deck i on trial T_{n+1}, given that participants begin the task with a state of no knowledge about the payoff structure of the decks. This expression is used in computation of the likelihood for the EVM, and a full derivation is provided in the supplementary material.

Third, the learned expectancies related to a deck will determine the probability of choosing any given deck (Busemeyer and Stout, 2002). The EVM represents this assumption by using a ratio-of-strengths rule to compare the probabilities of choosing each deck

$$Pr\left[D_{i,t+1}\right] = \frac{e^{E_{i,t} \times \theta(t)}}{\sum_{j=1}^{4} e^{E_{j,t} \times \theta(t)}} \tag{4.4}$$

where $E_{i,t}$ is the Expectancy for deck D_i on trial t as described in Equation 4.2. Alone, a ratio of strengths rule would ensure that if any deck has a much larger Expectancy than the remaining decks, then a choice from this deck will

be highly favoured (Busemeyer and Stout, 2002). However as participants become fatigued or bored as the task progresses, their choices may become less consistent than the ratio of strengths rule alone would explain. To allow for this, a trial dependant *sensitivity* function $\theta(t)$ is included in Equation 4.4, where

$$\theta(t) = \left(\frac{t}{10}\right)^c \tag{4.5}$$

for trial t with $-5 < c < 5$. Busemeyer and Stout (2002) argue that within Equation 4.5, negative values of c will represent boredom with the task or fatigue by decreasing the sensitivity to Equation 4.2 and increasing the number of random choices. As such, c is presented as representing *consistency* of choice. It should also be highlighted that c is included as a power of the sensitivity function and as such, very small changes in c result in large changes in the overall probability of choice.

Full descriptions of the model and its development are available elsewhere (Busemeyer and Stout, 2002; Wetzels et al., 2010).

4.3 Experiment 1

Despite being demonstrated in multiple studies to provide some prediction of risk of relapse, the standard measures of performance on the IGT (net return or frequency of deck selection) provide only composite measures of behaviour (Nejtek et al., 2013; De Wilde et al., 2013b; Radat et al., 2013; Goudriaan et al., 2011). Using the IGT as a clinical tool, with these composite measures of performance, provides an illustrative case where the evidence obtained may not be detailed enough to accurately inform decisions relating to treatment options (such as whether to continue current treatment, or to classify a person as having a high risk of relapse). In previous studies, the EVM has been used to disentangle these composite measures and compare the underlying psychological processes responsible for decision making

deficits on the IGT across groups of psychological interest (Peatfield et al., 2012; Brown et al., 2012; He et al., 2012; Yechiam et al., 2007; Bechara et al., 2000; Yechiam et al., 2008; Fridberg et al., 2010; Yechiam et al., 2005). Despite large amounts of uncertainty associated with individual level parameter estimates, the EVM of the IGT has been successful in differentiating groups of participants such as neurologically impaired from healthy individuals. By pooling information across multiple individuals, the uncertainty of group parameter estimates is reduced; however, if the EVM was to be used at the level of the individual instead, the uncertainty associated with EVM parameter estimates would have to be reduced at the level of the individual (Wetzels et al., 2010). With the aim of ascertaining why the observed parameter distributions are so highly variable, we consider a random effects model in which the random effect represents subject-to-subject variability. In addition, participants completed a 250 trial version of the task, rather than a 100 trial version as is often displayed in the literature. This is in line with the recent finding that 100 trials is insufficient for participants to learn the nature of the decks (Steingroever et al., 2013a).

The data provided for these analyses were collected from four drug use groups: MDMA (3.4-methylenedioxymethamphetamine) only consumers, cannabis only consumers, MDMA-and-cannabis polydrug consumers and drug-naïve controls. It has previously been shown that regular drug use is associated with increased levels of impulsivity (Quednow et al., 2007; Stout et al., 2005; Parrott et al., 2007; Yechiam et al., 2005) which, in the case of the IGT, may produce a strong focus on rewards and result in high attention to wins on the Valence parameter for the EVM. Higher outcome impulsivity is, therefore, expected in each of the drug using groups, manifesting in an increased positive affective reaction towards wins. It would be expected, therefore, that the values of W in the EVM will thus be increased.

The serotonergic impairment specifically associated with ecstasy use has also been shown to be related to learning and memory deficits and heightened impulsivity independent of the effects of use of other substances (Laws and Kokkalis, 2007; Evenden, 1999a). The MDMA use group would, therefore, be expected to display more volatile behaviour, and hence have higher values of ϕ than non-MDMA using groups since ϕ can be interpreted as an indicator of deficits on memory. Increases in ϕ are representative of a reduction in the ability to retain information over time (Lane et al., 2006; Stout et al., 2005).

4.3.1 Method

Participants

The data used for analyses were provided by the School of Psychology at the University of Tasmania. Information was provided for 25 participants from four different groups as follows:

1. Drug naïve controls (11 individuals with no history of drug use in the preceding six months and maximum lifetime history of five occasions of illicit drug use);

2. Regular users of MDMA (3 individuals with no history of cannabis use);

3. Regular Cannabis users (4 individuals with no history of ecstasy use);

4. Regular users of both MDMA and Cannabis (7 individuals).

No participant was in more than one group.

Materials

Each individual participated in a 250 trial version of the Iowa Gambling Task as created and described by Grasman and Wagenmakers (2005). The order in which the decks were presented were randomised for each task

and the order of wins and losses were randomised within ten trial blocks (Grasman and Wagenmakers, 2005). The current implementation made no changes to the defaults set in the program other than to define the number of trials as 250. That is, gain and loss amounts were not changed, the visual feedback function (which presents a happy or sad face when the individual makes a profitable or non-profitable choice respectively) was turned off, but graphical feedback was turned on. Graphical feedback consisted of informing the participant of how much they have won and how much they have lost on the current trial, the net amount they had before their most recent choice, their net total after their most recent choice, and the current trial number and the total number of trials.

4.3.2 Random Effects Models and Bayesian Estimation

In it's simplest form, Bayesian estimation makes use of Bayes formula:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where H is the hypothesis we are testing and E is the evidence collected. In this way we integrate our *prior* beliefs about the probability of the hypothesis, $P(H)$ with the likelihood of observing the collected data if the hypothesis were true, $P(E|H)$. After rescaling for the probability of observing the evidence independent of any hypothesis, $P(E)$, we gain an updated belief about the *posterior* probability of the hypothesis being true given the observed data, $P(H|E)$. In the current case, the "hypothesis" relates to the parameter estimates of the models under consideration and the priors will be defined by our beliefs surrounding their distribution. To begin with, however, the form of the likelihood must be defined.

For person k from group j on trial t, the standard EVM model is

$$y_{kjt} \sim \text{Multinomial}(1, \{p_{kj1}, p_{kj2}, \ldots, p_{kjt-1}\})$$

$$p_{kjt} = \text{EV}(t, W_{kj}, \phi_{kj}, c_{kj}, \{y_{kj1}, y_{kj2}, \ldots, y_{kjt-1}\})$$

where each individual has their own responses y_{kt} and probabilities p_{kt}. To incorporate random effects into the model, allowing for individual variability within estimates, for group j each of the EVM parameters W_{kj}, ϕ_{kj} and c_{kj} were constrained to be normally distributed

$$W_{kj} \sim \text{N}(\mu_{Wj}, \sigma^2_{Wj})$$

$$\phi_{kj} \sim \text{N}(\mu_{\phi j}, \sigma^2_{\phi j})$$

$$c_{kj} \sim \text{N}(\mu_{cj}, \sigma^2_{cj})$$

for independent W_{kj}, ϕ_{kj} and c_{kj}. This produced the classical likelihood for the full random effects model for each group

$$\begin{aligned} &L(y|\mu_W, \sigma^2_W, \mu_\phi, \sigma^2_\phi, \mu_c, \sigma^2_c) \\ &\quad = \prod_{kj} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(y_{kj}|W_{kj}, \phi_{kj}, c_{kj}) \\ &\quad \times P(W_{kj}|\mu_W, \sigma^2_W) P(\phi_{kj}|\mu_\phi, \sigma^2_\phi) P(c_{kj}|\mu_c, \sigma^2_c) \text{d}W \text{d}\phi \text{d}c \end{aligned} \tag{4.6}$$

where the three integrals represent multidimensional integrals over all W_{kj}, ϕ_{kj} and c_{kj} and with the probabilities P of the observed variables conditional on the random variables.

Equation 4.6 is analytically intractable and marginalizing over the random effects would require numerical quadrature (Pawitan, 2001). Due to the high degree of accuracy required and the computational intensity associated with numerical quadrature, this approach may not provide accurate

estimates. However, Bayesian estimation techniques avoid the problems associated with using Newton minimization in association with numerical quadrature (Congdon, 2003). In addition, for large data sets it is common to assume that parameter estimates are normally distributed and to estimate their covariances from the Fisher information (Pawitan, 2001). If these assumptions are not met, and for a finite sample there is no guarantee that they are, accurate point estimates and informative standard errors may be hard to obtain (discussed further in the supplementary material). The advantage of Bayesian estimation techniques are that they provide a mechanism for deriving the distribution of the parameters, rather than forcing the adoption of a normal approximation. It has been consistently shown that Bayesian estimation techniques result in more reliable estimates than maximum likelihood estimation (Ahn et al., 2011; Wetzels et al., 2010). For these reasons, this study uses Markov Chain Monte Carlo (MCMC) techniques to derive Bayesian estimates of parameters in the EVM.

4.3.3 Markov Chain Monte Carlo Sampling

The benefit of using Markov Chain Monte Carlo (MCMC) methods is that the complex search for maximum likelihood solutions to multi-parameter problems can be avoided by sampling directly from the posterior density (Congdon, 2003). Analysis was completed using a probit transformation. The probit transformation negates the need for truncated normal priors on bounded parameters as it transforms a (0,1) constraint on a parameter to (-∞, ∞). Once analysis was completed, results were back-transformed into their original parameter domains.

The full posterior is proportional to

$$P(\mu_W, \mu_\phi, \mu_c, \tau_W, \tau\phi, \tau_c|y_{kt}) \propto P(y|\mu_W, \mu_\phi, \mu_c, \tau_W, \tau\phi, \tau_c) \times P(\mu_W)P(\mu_\phi)P(\mu_c)P(\tau_W)P(\tau_\phi)P(\tau_c). \quad (4.7)$$

where we have parametrized in terms of the precisions $\tau_W = \sigma_W^{-2}$, $\tau_\phi = \sigma_\phi^{-2}$ and $\tau_c = \sigma_c^{-2}$. The first term on the right hand side of Equation 4.7 is the likelihood described in Equation 4.6 and the remaining terms represent the prior distributions. Independent, diffuse priors were assumed for the model parameters.

The Metropolis sampler (Gilks et al., 1995) with a multivariate normal proposal distribution was used to sample W_{kj}, ϕ_{kj}, c_{kj} for each individual in each group, conditional on the $\mu_W, \mu_\phi, \mu_c, \tau_w, \tau_\phi, \tau_c$. A Gibbs sampler was used to sample for the population parameters $\mu_w, \mu_\phi, \mu_c, \tau_w, \tau_\phi, \tau_c$ conditional on the individual random effects $W_{kj}, \phi_{kj}, c_{kj}$.

In total, 25,000 MCMC samples were drawn from the posterior distribution of each parameter for each individual. The MCMC samples were thinned by retaining only every tenth sample to reduce autocorrelation, and the first 500 retained samples were discarded to reduce dependence on the starting values. Analysis was completed using the statistical program R (The R Development Core Team, 2009), and the code for all computations is available by contacting the author directly.

Exploratory analysis of posterior distributions of the parameters in the EVM revealed cases for which the posterior was bi-modal. When a distribution does not have connected support, it can become hard for a single chain to sample the entire posterior and it can become trapped near one mode (Gilks et al., 1995). If this happens, convergence statistics can be good, but only because the chain is not aware of (or has not sampled) the remaining mode or modes. Using a multiple chain MCMC sampling method can, therefore, provide a more even and representative sample from the target distribution than using a single chain alone (Chib and Greenberg, 1995). We thus used multiple chain MCMC techniques running five chains from random starting values for each individual.

Raftery-Lewis diagnostics were completed to ascertain the effective sample size (ESS) required to achieve convergence of the chains and to ascertain if the chains were too highly autocorrelated or if they were 'sticking' in modes (Raftery and Lewis, 1992). Convergence of the chains was tested using Gelman and Rubin's multivariate convergence diagnostic, the multivariate potential scale reduction factor (MPSRF) (Brooks and Gelman, 1998).

To test the accuracy of our implementation of the model, IGT data was simulated for 95 synthetic participants across a full range of values for the parameters of the EVM. The EVM was then fit to the generated data using the formulas described above to see if the parameter values could be recovered. Posterior predictive checks were also completed by drawing 1,000 posterior samples per participant, simulating a new data set for each sample and then comparing the net return of the simulated data with the net return of the observed data. If the observed net return was within a 95% for the mean of the simulated net returns, this was deemed a successful fit.

4.3.4 Results

For a psychological assessment tool to be useful, it needs to be universally effective across the populations it is designed for. The following results have been chosen as examples of why the described analyses of performance on the IGT are not universally effective. If enough examples can be found showing a tool does not work, or that the interpretation of performance on the tool is inaccurate, then the applicability of the tool and/or the measurement of performance on the task must be brought into question. In the current study, use of the EVM to analyse performance on the IGT produced ineffective and potentially inaccurate results more often than it produced reasonable results. The individual results presented are good examples of the described

phenomena but they are also representative of the entire sample. The deck selection profiles and posterior distributions of the parameter estimates for the EVM for all participants are presented in the supplementary material.

Frequency of Deck Choices and Net Return

Several studies have compared relative frequencies of deck choices as a measure of performance on the IGT, suggesting that similar frequency of deck choices reflects similar decision making behaviour, implying similar degrees of underlying neurological functioning (Poletti et al., 2011; Brown et al., 2012; Lamers et al., 2006; Stout et al., 2004; Ahn et al., 2008; Yechiam and Ert, 2007). Figure 4.1 shows the overall frequencies of deck choice across the four groups (Group 1 - Drug naïve, Group 2 - MDMA only, Group 3 - Cannabis only, Group 4 - MDMA and Cannabis). With the possible exception of the ecstasy group, there is little difference in the overall frequency of deck selection across the four groups, suggesting that differences in behaviour are not represented by the overall frequency of deck selection, but in the sequence with which decks are selected. Consider Figures 4.2 and 4.3 which show two regular users of both cannabis and ecstasy with a similar frequency of deck choices. Despite the similarity in in frequency of deck choice, the net return experienced is quite different, with net return displayed in Figure 4.2 $2550 less than that displayed in Figure 4.3.

Due to the randomness with which win and loss amounts are ordered when completing the task, although participants may have similar frequency of deck choice, the returns they experience on any of those choices will vary and, as such, each participant will have a unique experience of the deck returns. It is the way the individual responds to this unique experience that defines an individual's overall performance, not the number of times the individual chooses a deck. This supports previous findings that suggest

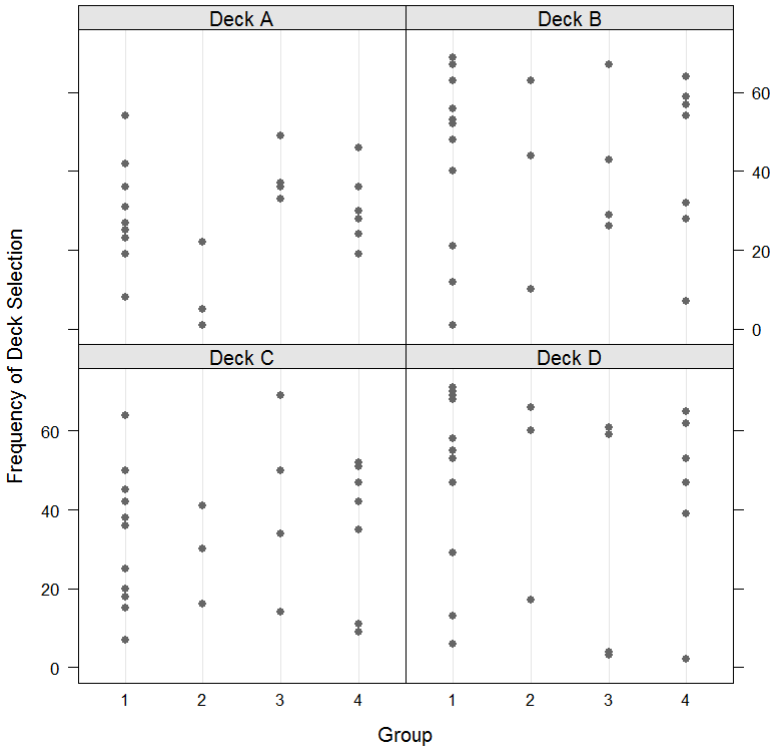

Figure 4.1: Frequency of deck selections for each deck by participant group. In this figure, group 1 is the drug naïve group, group 2 are regular users of ecstasy, group 3 are regular users of cannabis and group 4 are regular users of both ecstasy and cannabis. Each dot represents the frequency of deck choice for an individual participant. A large amount of variability in frequency of deck choices within each group is visible with little discernible differences apparent between groups.

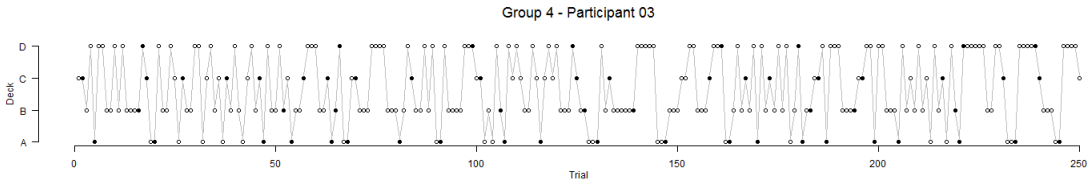

Figure 4.2: Deck selection profile of IGT performance for a regular users of both cannabis and ecstasy. A circle represents a choice from the corresponding deck label on the left of the figure and if the circle is filled, then a loss was experienced. The net return experienced for this participant was an overall loss of $650 with Decks A chosen approximately 16%, B 34%, C 16% and D 33% of the time. Comparing this profile to the profile displayed in Figure 4.3 shows that despite having similar frequency of deck choices, the net return acquired by the participant in Figure 4.3 (an overall win of $1900) is extremely different to this participant.

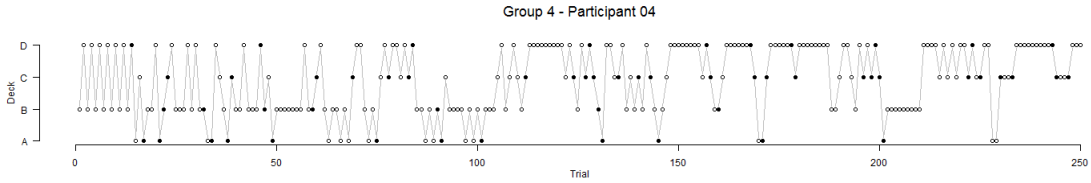

Figure 4.3: Deck selection profile of IGT performance for a regular users of both cannabis and ecstasy. A circle represents a choice from the corresponding deck label on the left of the figure and if the circle is filled, then a loss was experienced. The net return experienced for this participant was an overall win of $1900 with Decks A chosen approximately 10%, B 30%, C 20% and D 40% of the time. Comparing this profile to the profile displayed in Figure 4.2 shows that despite having similar frequency of deck choices, the net return acquired by the participant in Figure 4.2 (an overall loss of $650) is extremely different to this participant.

collapsing choices across the task into good and bad deck categories does not capture performance well (Steingroever et al., 2013a).

Figures 4.4 and 4.5 show two participants with similar net return ($2950 and $3050 respectively) but choice strategies which are quite different. Although Figure 4.5 shows a net return of only $100 more than that shown in 4.4, unlike in Figure 4.4 this participant clearly did not converge to a preferred deck and was still making somewhat erratic choices throughout the duration of the task. Again, this would support the idea that net return is not capturing the complexity of the decision making process required to complete the IGT. Using frequency of deck choice or net return to analyse performance on the IGT is, therefore, not recommended and will not be pursued further in this chapter.

Parameter Recovery and Posterior Predictive Checks

From the 95 synthetic participants (with known values of W, ϕ and c) the credible intervals of the posteriors contained the imputed parameter value of $W \simeq 97\%$ and $\phi \simeq 94\%$ of the time. This suggests that the Bayesian estimation of the EVM described here is able to recover the imputed parameter values of W and ϕ with reasonable accuracy. As previously described, very

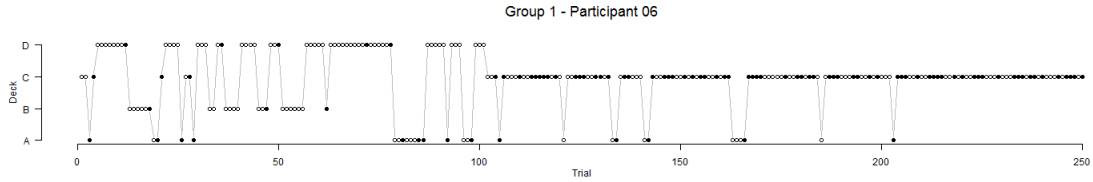

Figure 4.4: Deck selection profiles of IGT performance for a healthy drug naïve participant displaying convergence of deck choices to favoured decks. A circle represents a choice from the corresponding deck label on the left of the figure and if the circle is filled, then a loss was experienced. The net return experienced for this participant was an overall win of $2950 with Decks A chosen approximately 12%, B 9%, C 57% and D 22% of the time. Despite the similarities in net return between this participant and the participant displayed in Figure 4.5 (with an overall win of $3050), the strategies of deck choice are not similar as in Figure 4.5 the participant continues to make less consistent choices throughout the task.

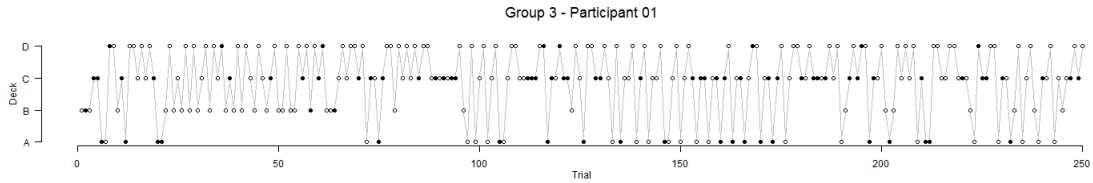

Figure 4.5: Deck selection profiles of IGT performance for a regular user of cannabis displaying inconsistent choices throughout the task. A circle represents a choice from the corresponding deck label on the left of the figure and if the circle is filled, then a loss was experienced. The net return experienced for this participant was an overall win of $3050 with Decks A chosen approximately 16%, B 13%, C 39% and D 32% of the time. Despite the similarities in net return between this participant and the participant displayed in Figure 4.4 (with an overall win of $2950), the strategies of deck choice are not similar with this participant failing to converge to clearly preferred decks as the participant displayed in Figure 4.4 has done.

small changes in c result in large changes to the probability of deck choice, effectively reducing the observable range of values of c to something much narrower than -5 to 5. For reasonable imputed values of c (between 0.3 and 0.7) the imputed value was recovered approximately 67% of the time. This suggests that although the Bayesian estimation of the EVM described here is able to recover the imputed parameter values with reasonable accuracy, there may be some elements of the EVM system of equations that make it impossible to recover behaviour accurately.

Further to this, posterior predictive checks were completed for each observed participant. Although the net return for each participant was generally within an expected 95% credible interval for the net return, the span of expected net returns covered by the credible intervals were quite large (approximately \$9000, see Figure 4.6). This finding suggests that there is a large amount of uncertainty associated with estimates gained using the EVM and, supportive of the findings from the synthetic participants, that it may be impossible to recover behaviour accurately. The supplementary material has further details of the posterior predictive checks.

Participant Based Posterior Results

Of the 25 real participants, only 2 had approximately normally distributed and reasonably localised parameter estimates. Of the remaining 69 parameter estimates (23 participants), 18 were bi-modal, 12 were on the boundary of parameter space, 7 spanned the entire parameter range and 18 non-linear relationships between parameter estimates were observed. The posteriors, with trace plots to display the mixing of the chains, for all participants are presented in the supplementary material. Raftery's dependence factor, I, was close to 1 for all parameters for each participant suggesting that autocorrelation between posterior samples was not a problem and that the chains

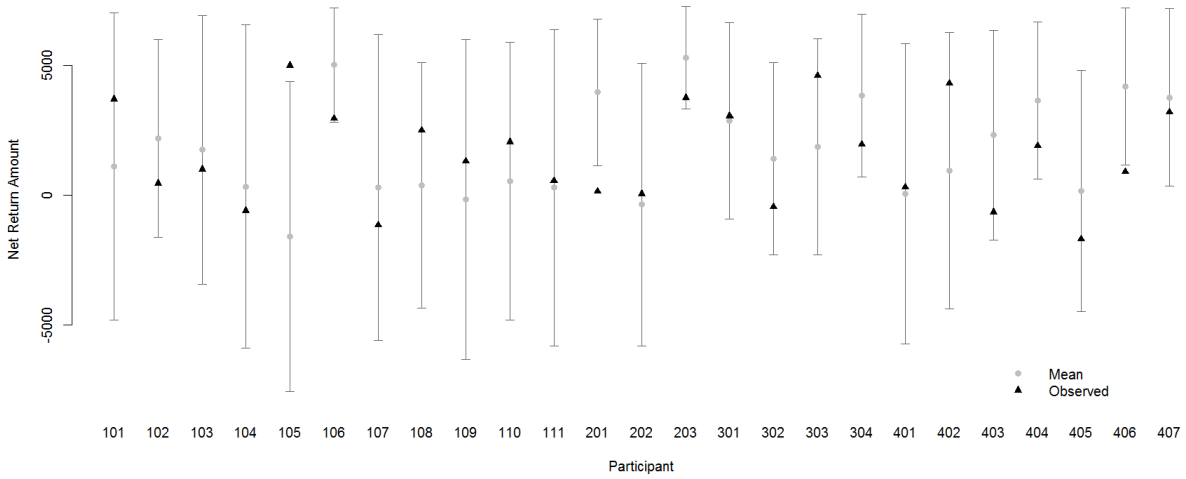

Figure 4.6: 95% Credible intervals for the net returns for each of the 25 participants, generated from posterior predictive checks. It is clear that although all but three credible intervals contain the observed net return, the variability of the credible intervals is extremely large. This would suggest that there is uncertainty surrounding the predictions from the model.

had mixed well. In addition, all MPSRF values were less than or equal to 1.1 for all but one participant, confirming convergence of the chains. The MPSRF for the remaining participant did not reduce below 4.3, even when 8 chains with 120,000 MCMC samples each were drawn from the posterior distribution (thinning of 10, burn-in at 2,000). Figure 4.7 shows this participant, from the ecstasy-and-cannabis use group, has a bi-modal distribution of estimates for parameters ϕ and c with no connected support between the modes. As discussed in Section 4.3.3, when there are two or more disconnected modes, it becomes harder for a Markov chain to sample all of the posterior as the chain can become 'stuck' in one of the modes. If this happens, convergence statistics can be good, but only because the chain is not aware of (or has not sampled) the remaining mode or modes. Figure 4.8 shows the mixing for the posterior in Figure 4.7 for the five separate chains used in this analysis. This shows that two of the five chains have sampled one mode, while the remaining three have sampled the alternate mode, for the duration of the chain. This lends weight to the argument that

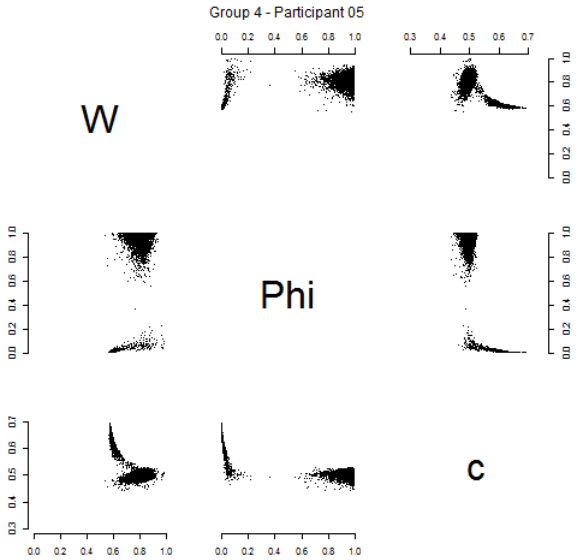

Figure 4.7: The sampled posterior for a participant from the ecstasy-and-cannabis use group using the Expectancy Valence Model. This reveals a bi-modal distribution in ϕ and W.

the posterior may not have converged due to posterior distributions being truly bi-modal. Bi-modal distributions such as that in Figure 4.7 cannot be approximated well by a point estimate thus an individual's behaviour is best summarised by considering the full posterior distribution.

For 12 of the 25 participants, analysis using the EVM produced posterior densities on the boundaries of W and/or ϕ. For example, Figure 4.9 shows that, for this participant, the density for ϕ is hard against the upper boundary value of one (observable in the middle of the top and bottom rows and the left and right columns of Figure 4.9). This is in line with the findings presented by Wetzels et al. (2010) who, due to problems with maximum likelihood estimation, excluded approximately 25% of each of their four data sets due to estimates on the boundary. Rather than excluding these observations, using the posterior mean with a credible interval here, would provide a more reasonable estimate than an MLE (supplementary material).

Also note the strong, non-linear relationship between parameters ϕ and c in Figure 4.10. Although statistically non-linear relationships of this type

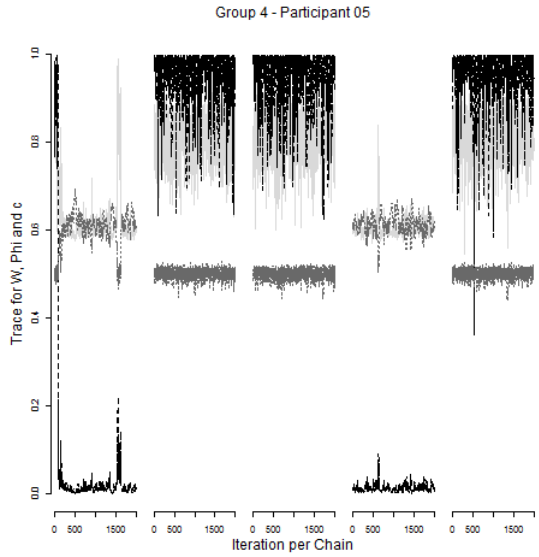

Figure 4.8: The mixing of the posterior distribution displayed in Figure 4.7 for ϕ (black), W (light gray) and c (dark gray). Each block of 2000 iterations represents one of the five thinned chains once burn-in was discarded. Two distinct solutions are displayed, corresponding to the bi-modality observed in Figure 4.7. Each of the five chains has sampled to one of the two modes.

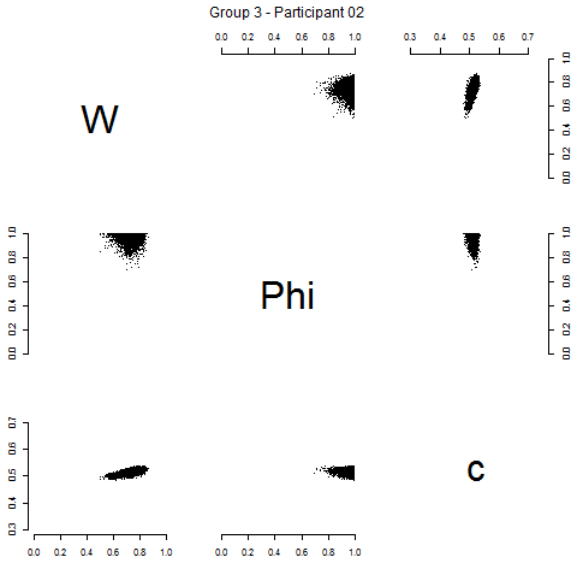

Figure 4.9: The sampled posterior distribution for a participant from the regular cannabis use group using the Expectancy Valence Model. The posterior distribution provides evidence of estimates with distributions on the boundary of parameter space.

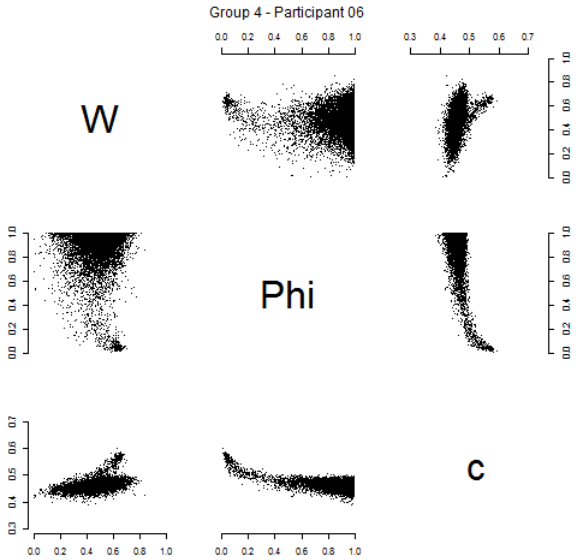

Figure 4.10: The sampled posterior distribution for a participant from the ecstasy-and-cannabis use group using the Expectancy Valence Model. The posterior distribution provides evidence of non-linear relationships between estimates.

can be transformed and interpreted mathematically, psychologically any interpretation of this sort may make little or no sense, as we elaborate in the next section.

Psychological Inference

Results suggest that individual performance on the IGT can be mathematically described by looking at the full posterior distribution. However, the mathematical descriptions may not be psychologically meaningful.

For the participant in Figure 4.7, the distribution of the parameter c suggests that deck choices were either perfectly consistent or highly inconsistent across trials and the distribution for ϕ suggests that the decisions made were either highly influenced by recent outcomes, or were influenced solely by the initial expectancies, or a state of no knowledge. Interpreting the relationship between ϕ and c suggests that the participant was either inconsistent when memory was depleted or varied in their level of consistency when memory of outcomes was good. So for this participant, it is unclear whether he or she has a deficit in memory or not. In general, for

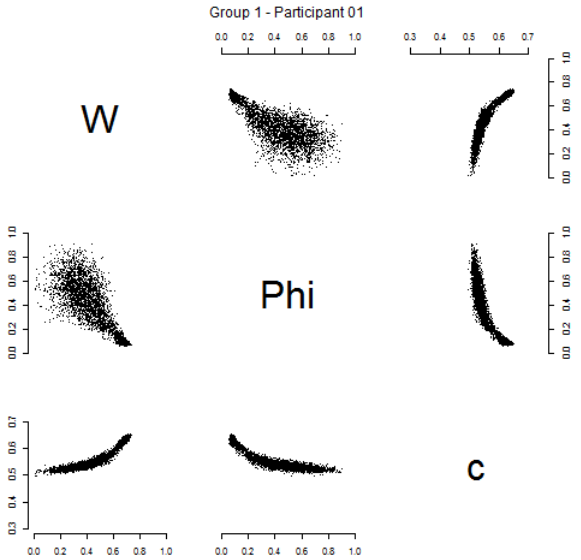

Figure 4.11: The sampled posterior distribution of a drug-naïve participant using the Expectancy Valence Model. Distinct non-linear relationships between the posterior distributions of ϕ and c and between W and c are visible in the bottom row and right column. There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.12, 4.13 and 4.14. This is despite the fact that each of these participants come from the same group of psychological interest.

participants with bi- or multi-modal distributions, conflicting psychological inferences may be drawn about deficits in decision making, undermining the application of the EVM.

Non-linear relationships between parameters also produce uncertainty surrounding the observed relationship between the model and the measured variables which can lead to competing and conflicting psychological interpretations for individuals. The estimates provided in Figure 4.11 suggest that the participant is either a consistent but impulsive individual with good memory or an individual with poor memory who is inconsistent in their choices and focuses more on losses than wins. So it is unclear where any deficits in decision making may lie for this individual.

For the individual depicted in Figure 4.12 it appears that memory for recent outcomes varies across the entire range of values for ϕ, similarly for affective reaction to deck outcomes W. Psychologically, the goal is to

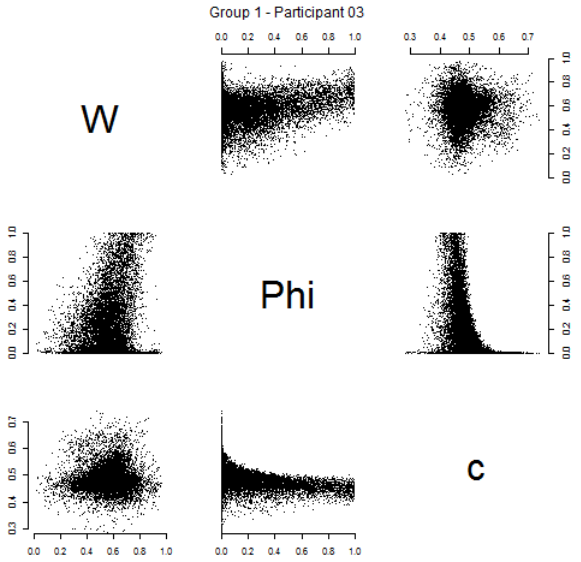

Figure 4.12: The sampled posterior distribution of a drug-naïve participant using the Expectancy Valence Model. The posterior distribution for ϕ here spans the entire parameter space (from $\phi = 0$ through to $\phi = 1$). There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.11, 4.13 and 4.14. This is despite the fact that each of these participants come from the same group of psychological interest.

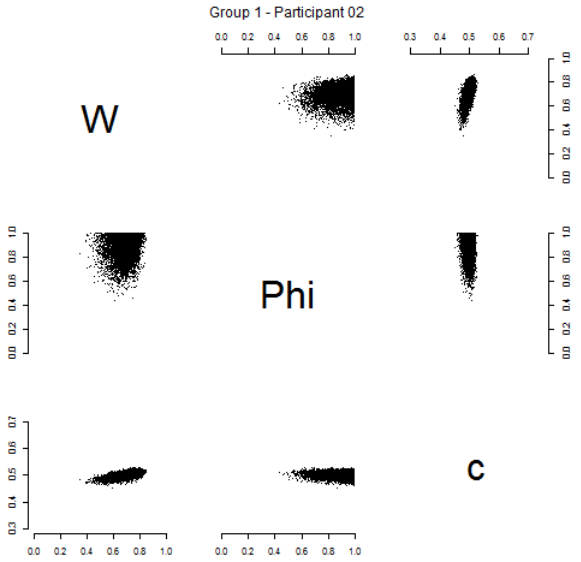

Figure 4.13: The sampled posterior distribution of a drug-naïve participant using the Expectancy Valence Model. Both the middle row and middle column show the posterior density for ϕ is hard against the upper boundary of parameter space (at $\phi = 1$). There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.11, 4.12 and 4.14. This is despite the fact that each of these participants come from the same group of psychological interest.

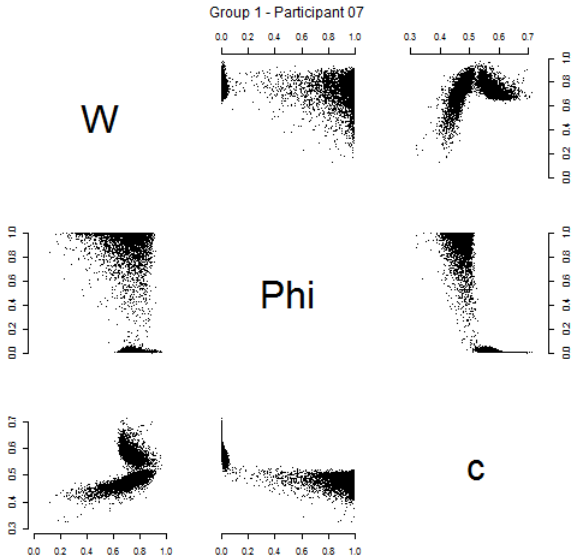

Figure 4.14: The sampled posterior distribution of a drug-naïve participant using the Expectancy Valence Model. Bi-modality of the sampled posterior distributions for both ϕ and c are evident. There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.11, 4.12 and 4.13. This is despite the fact that each of these participants come from the same group of psychological interest.

make statements about levels of impulsivity or memory for this individual. However the results in Figure 14.12 suggest that the participant may be highly impulsive, have normal levels of impulsivity or be extremely cautious. He or she may also have no memory for recent outcomes, have some memory impairment or have perfect memory. With such results it is again unclear whether the participant has displayed deficits in decision making processes or not.

Psychologically then, the EVM provides neither clear nor consistent estimates of neurological process that could be used to gain specific insight into an individual's decision making abilities. Instead, findings using the EVM at the individual level seem confusing, conflicting and hard to interpret.

Group Estimates

We observed a lack of consistency within the groups of psychological interest that were tested. Figures 4.11 to 4.14 show the sampled posterior distributions for participants from the drug naïve group who should not display any

particular deficits in the memory parameter ϕ (Quednow et al., 2007; Stout et al., 2005; Parrott et al., 2007; Yechiam et al., 2005). Figure 4.13 shows values of ϕ indicative of a short associative memory, while Figure 4.11 shows values indicative of a long associative memory. Alternatively Figures 4.12 and 4.14 reveal distributions suggesting both high and low values of ϕ. So, a point estimate of group performance would also carry a large associated variance reflecting the inconsistency of individual-level parameter estimates within the group.

Although parameter estimates obtained using the EVM across groups have been shown to be effective in identifying differences between groups of individuals (for example Yechiam et al. (2005)), the question is now whether it is appropriate to do so. Figure 4.15 shows the EVM parameter estimates for the drug naïve control group which suggests that the drug naïve controls had long associative memories (ϕ close to 0) and this is supported by the 95% credible interval for the parameter estimates displayed in Table 4.2. However we have already seen in Figures 4.12, 4.13 and 4.14 that this is not necessarily the case for individual participants. In fact, out of the 11 participants in the drug naïve control group, only 2 had clear estimates of ϕ that were close to 0. Of the remaining 9 participants, 4 had estimates of ϕ that were bi-modal, 3 spanned the entire (or close to entire) parameter space and 2 had estimates of ϕ that were the complete opposite to the group estimate. Psychologically then, the group estimate suggests long associative memories for drug naïve controls, but individual estimates suggest that up to 9 out of the 11 drug naïve controls may have deficits in memory processes.

If a group is made up of individuals with parameter estimates that are bi-modal, non-linear, span every possible value and are inconsistent across the group then the meaning of any group estimate is questionable. Due to this variability in parameter estimates for participants in the same group,

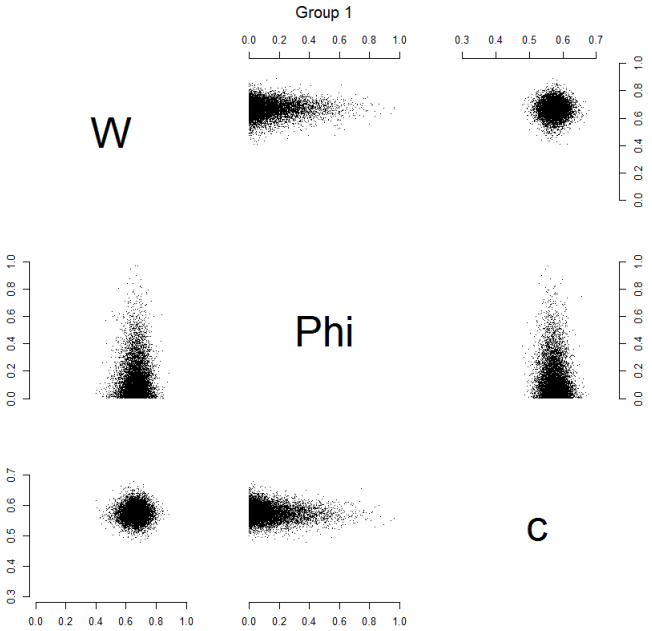

Figure 4.15: The sampled posterior distribution of the drug naïve group using the Expectancy Valence Model. Here a value of ϕ close to zero is suggested, however this is in direct contrast to some of the participants who make up the group (for example see Figure 4.13) that have estimates of ϕ close to one.

Table 4.2: 95% credible intervals for W, ϕ and c each of the four drug using groups.

	95% credible Interval					
	W		ϕ		c	
Group	Lower	Upper	Lower	Upper	Lower	Upper
Drug Naïve	0.56	0.77	0.00	0.39	0.53	0.62
MDMA	0.24	0.69	0.10	0.69	0.44	0.60
Cannabis	0.36	0.73	0.00	0.94	0.49	0.63
MDMA & Cannabis	0.36	0.70	0.00	0.32	0.46	0.59
Overall	0.33	0.76	0.00	1.00	0.46	0.62

any point estimate for the group may not be a true representation of this diversity of psychological outcomes.

4.3.5 Summary

Bayesian analysis of the psychological processes underlying performance on the IGT using the EVM did not produce localized, uni-modal or approximately normal estimates of parameter distributions. When considering the full posterior distributions, the psychological interpretation of the results

was contradictory, uninformative and, at times, misleading. This suggests that there is no guarantee, when applied at the level of the individual, that point estimates of EVM parameters are accurately measuring individual deficits in decision making.

Using the EVM to estimate the psychological processes leading to decision making deficits in group performance may produce reasonably localised estimates. However with individuals within the group receiving estimates that have multiple conflicting psychological interpretations, the interpretation of any group estimate may not be truly representative of the group. Obtaining clearer, more localised estimates for each individual may help to reduce this variability.

4.4 Experiment 2

Measuring a single run (one run of 250 trials) on the IGT, as presented in Experiment 1, produces a single observation for each participant. We increased the number of runs each participant completed to ascertain if less variable individual-level parameter estimates for an individual could be obtained when decomposing behaviour on the IGT using the EVM.

Participants were asked to complete three runs (three runs of 100 trials each) of the IGT. Even though this only provides 50 more individual choices than in the Experiment 1, a participant must start from a state of no knowledge three times in this experiment compared to only once in the first. Given that the same psychological processes are responsible for the performance on each of the three runs, instead of gaining one measure of initial learning, there are now three measures per participant. As such, the sample size for each participant is increased from one to three. For each participant then, multiple sets of responses can now be used to estimate the overall performance on the IGT for that participant. It would be expected

that posterior distributions gained in this way would be less variable than those observed in Experiment 1.

As with Experiment 1, we consider a random effects model of multiple runs per participant to account for subject to subject variability.

4.4.1 Method

Participants

Eight participants were recruited using posters in the Psychology Department at the University of Tasmania and consisted of four males and four females, with mean ages of 37.5 ($SD = 14.9$) and 45.5 ($SD = 17.1$) respectively. Any person who was over the age of 18 and reported no cognitive deficits was able to participate and as such, this group was considered a random group of healthy controls. Adherence to the participation criteria was ascertained by self-report. Each participant gave informed consent and was able to withdraw at any time without enquiry.

Materials

The Iowa Gambling Task (IGT) was implemented using the computerized test described by Grasman and Wagenmakers (2005). The order of presentation of the four decks in the IGT was random, so a participant completing three sequential runs would be unlikely to gain the same ordering of the decks (Grasman and Wagenmakers, 2005). Participants were made aware of the fact that the decks would be different each time they completed the task. Loss amounts are also randomly presented such that the required ratio of wins to losses are reached for each deck without a predictable pattern (Grasman and Wagenmakers, 2005). Our only modification to the task was to restrict the number of trials to 100 for the main task and to include 10 trials for a practice task.

To account for learning effects between runs, distractor tasks were em-

ployed.

Procedure

After an initial practice run of 10 trials, participants completed the first of three 100 trial runs after which two distractor questionnaires were completed. The second 100 trial run was then completed after which participants completed a computerized cognitive test (a Stroop task). A third and final run of 100 trials was then completed.

The distractor tasks employed here were used to apply some cognitive load such that participants could not review their performance on the task and make plans for future runs. If the distractor tasks were not successful in achieving this, there should be sequential improvements across the runs for the participants.

4.4.2 Results

Analysis of the psychological processes underlying performance on the IGT was completed with the EVM and estimation of the posterior distribution for each parameter was averaged across performance in all three runs. The following results have been chosen as good examples of why the described analyses of performance on the IGT are not universally effective. The individual results presented are good examples of the described phenomena and they are loosely representative of the entire sample. Deck selection profiles of the three runs for each participant are provided in the supplementary material along with posterior distributions of the parameter estimates for the EVM for all eight participants completing the IGT.

Net Return

Initial exploration of raw data showed that for each participant the net amount won at the end of each run varied considerably (Figure 4.16). In fact, Figure 4.16 shows that half of the eight participants varied between gross

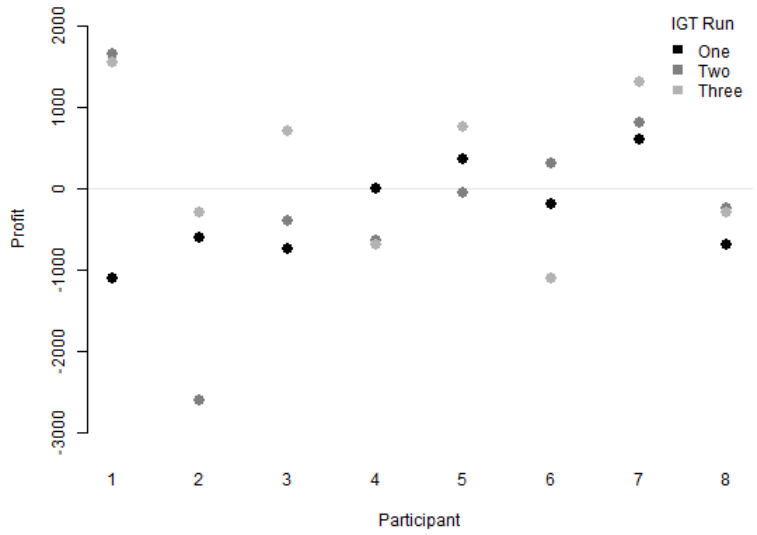

Figure 4.16: The profit for each of three runs of the IGT completed by the eight participants in Experiment 2. The large amount of variability between runs on the IGT for each participant is apparent here with five participants showing greater than $1400 between their smallest and largest profits.

profits and losses between runs and five of the eight participants showed differences of greater than $1400 between their best and worst performances. There was no consistent improvement across the runs, suggesting the distractor tasks accounted for learning effects, and instead, on any given run, performance in terms of a net outcome varied depending on the pattern of observed values related to deck choices. Participant 2, for example, had net outcomes of $1400 on run 1, -$600 on run 2 and $1700 on run 3. If taken individually, a net outcome of $1400 or $1700 would suggest good performance, while a net return of -$600 would indicate poor performance and neurological deficits in decision making.

The lack of consistency in the outcomes achieved on the IGT raises some questions about the reliability of the task as a clinical tool. In fact, reviews of the IGT by both Buelow and Suhr (2009) and more recently Lin et al. (2013) highlight the need for better information about the reliability of the task. We would expect, if the IGT were reliable, that participants should achieve similar results each time they complete the task, but this is

clearly not the case here. Further, if the differences in net return observed in Figure 4.16 were due to systematic carry-over effects (i.e. learning) we would expect to see participants getting better over time. Figure 4.16 shows this is also not the case. Therefore, this may suggest an inherent problem with the reliability of the IGT itself and further investigation into this would be worthwhile.

Participant Based Posterior Results and Psychological Inference

Of the eight participants, two had posterior distributions that were bi-modal in ϕ and c (Figures 4.17 and 4.18), three exhibited clear non-linear relationships between ϕ and c, ϕ and W and/or W and c (Figure 4.17, 4.19 and 4.20) and one showed a posterior distribution spanning the entire parameter space for ϕ (Figure 4.20). Convergence of the MCMC procedure was assured with the MPSRF (Brooks and Gelman, 1998) reading below 1.1 for every participant.

Psychological interpretation of the observed posteriors, much like in Experiment 1, was still unclear. Participant 1, for example, appears to be highly impulsive (W) with perfect memory (ϕ) and moderate consistency of choice (c) but also may have deficits in memory and have a tendency toward focussing on wins (Figure 4.18). Here analysis of the posterior distribution is possible but, due to contradictory results from bi-modal and non-linear distributions, not psychologically meaningful.

Group Estimates

Figures 4.17 to 4.20 show a lack of consistency between participants in the group, which is unexpected for a group of healthy controls (although it should be noted that a similar result was apparent in the healthy control group for Experiment 1). Although the bi-modality and non-linearity displayed in Figures 4.17 to 4.20 is not as severe as that displayed in Experiment

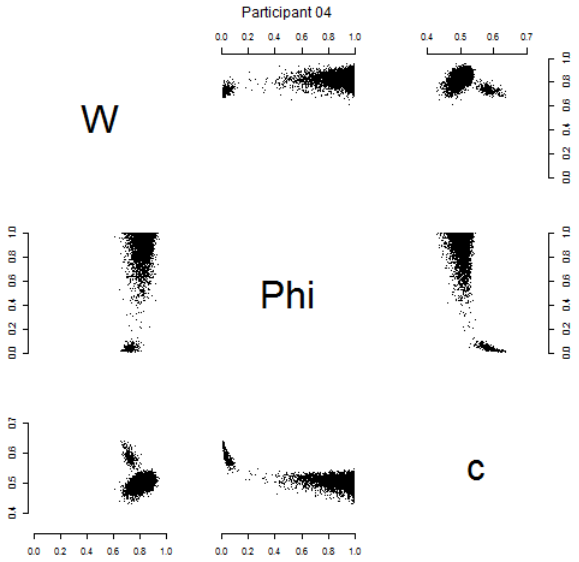

Figure 4.17: The sampled posterior distributions for participant 4 using the Expectancy Valence Model to decompose behaviour on a multiple run version of the IGT. This figure shows bi-modal distributions in ϕ and c similar to those observed in Figure 4.6 which was based on a single run of the IGT. There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.18, 4.19 and 4.20. This is despite the fact that each of these participants come from the same group of psychological interest.

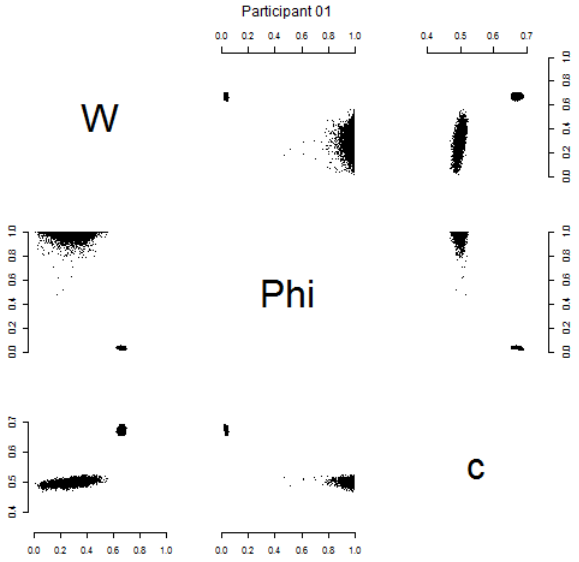

Figure 4.18: The sampled posterior distributions for participant 1 using the Expectancy Valence Model to decompose behaviour on a multiple run version of the IGT. The posterior samples here show estimates of ϕ hard against the upper boundary of 1 similar to those observed in Figure 4.8 which was based on a single run of the IGT. However, here we also see distinct bi-modality in posterior estimates for both ϕ and c. There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.17, 4.19 and 4.20. This is despite the fact that each of these participants come from the same group of psychological interest.

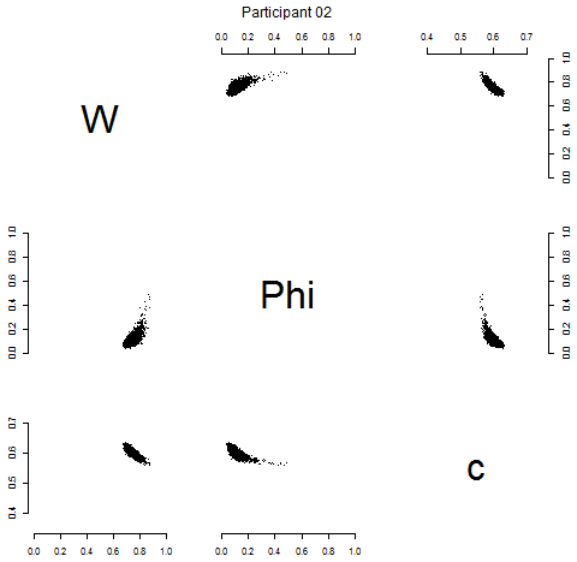

Figure 4.19: The sampled posterior distributions for participant 2 using the Expectancy Valence Model to decompose behaviour on a multiple run version of the IGT. This shows estimates of the parameter c are non-linearly related to ϕ. There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.17, 4.18 and 4.20. This is despite the fact that each of these participants come from the same group of psychological interest.

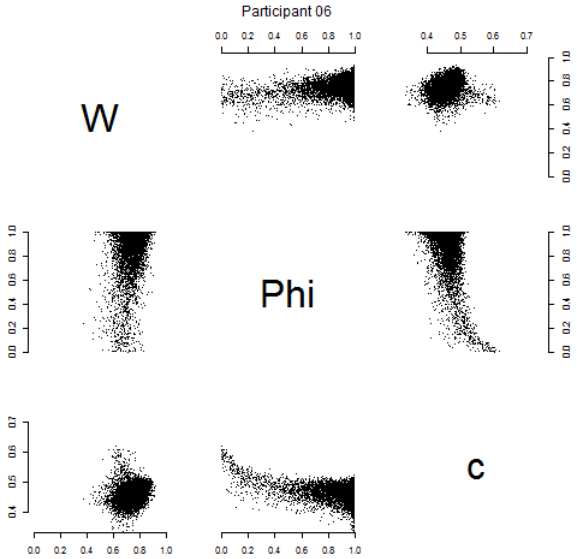

Figure 4.20: The sampled posterior distributions for participant 6 using the Expectancy Valence Model to decompose behaviour on a multiple run version of the IGT. This figure shows a posterior distribution for ϕ that takes all possible values from 0 to 1. This is similar to that observed in Figure 4.9 which was based on a single run of the IGT. We also see here a non-linear relationship between the estimates for ϕ and c. There is also a lack of consistency between the results displayed in this Figure and the results displayed in Figures 4.17, 4.18 and 4.19. This is despite the fact that each of these participants come from the same group of psychological interest.

1, the variability within groups is still high. This suggests that the issues associated with using the EVM that were identified in Section 4.3.4 still apply when the IGT is completed three times.

4.4.3 Summary

In summary, the hypothesis that there should be decreased uncertainty in individual-level EVM parameter estimates when participants completed the IGT three times compared to results from the single run IGT was not supported. Despite a reduction in individual uncertainty, the posterior distributions were not normally distributed and, due to bi-modality and non-linearity of estimates, produced conflicting psychological interpretations.

In Experiment 2 participants completed the IGT three times. An increase to five repetitions of the IGT was considered, but due to the lack of improvement over the three runs, this was deemed unnecessary. From a logistical standpoint increasing the number of runs of the IGT that an individual completes, to the point where the increased sample size noticeably decreases the uncertainty in individual-level parameter estimates, is not feasible. Increasing the number of runs would make completion of the IGT laborious for participants and lessen the likelihood of its use as a diagnostic tool.

Inconsistency between participants in the same group, as with Experiment 1, was also observed. Given the multiple, conflicting interpretations observed at the level of the individual, it is questionable whether a simple group estimate would truly represent the diversity of the group.

Additionally there was a lack of consistency between the net outcomes achieved by each participant across the three runs of the IGT. This inconsistency could not be explained by carry-over effects and as such raises questions about the reliability of the IGT itself. Agreeing with the findings

from Experiment 1, these findings suggest that the reliability of the IGT is something that would benefit from further investigation.

4.5 Experiment 1 Revisited

The EVM, when applied to the IGT, aims to identify the underlying psychological processes of memory and consistency of choice that lead to deficits in decision making. The multiple conflicting parameter estimates displayed in Experiments 1 and 2, however, suggest that the EVM is not always able to distinguish between different explanations of behaviour. For example if a participant has great memory but is bored and making random choices they may appear very similar to a participant who is trying really hard but has terrible memory. These two situations could lead to the same observable behaviour on the IGT and, with the same observed behaviour, it would be impossible to tell the two situations apart.

In a study by van Ravenzwaaij et al. (2011), who looked at modelling behaviour on the Balloon Analogue Risk Task (BART), the inability to accurately identify parameters was encountered when using the four parameter BART model proposed by Wallsten et al. (2005). van Ravenzwaaij et al. (2011) found that two of the four parameters were not identifiable and that a reduced, two parameter model provided the best fit. Given the multiple, conflicting parameter estimates of the EVM parameters when applied to IGT data, the following experiment explores whether a reduction in the number of parameters of the EVM may provide more accurate estimates for the remaining parameters of the EVM.

In psychological assessment, careful standards are followed to ensure consistency of test environment and optimal motivation of participants throughout testing (Kaplan and Saccuzzo, 2012). Outside of forensic contexts, psychologists generally assume optimal effort and motivation throughout test-

ing, and psychological literature either makes this implicit assumption by not including tests of effort or apply methods to temper the effects of inattention or lack of understanding. However over a short duration task it is reasonable to assume maximal effort and, as such, that participants have maintained consistent effort during the relatively short duration of the IGT presented here. Assuming consistent effort means that the consistency parameter in the EVM, c, would no longer be required to take into account boredom or fatigue and could be held constant.

In addition, the simulations presented in Section 4.3.4 show that, although parameters W and ϕ were recovered with reasonable precision 97% and 94% of the time respectively, the consistency parameter c was only recovered 67% of the time despite the fact that c was only simulated within a restricted band of $0.3 < c < 0.7$. This adds further weight to the suggestion that a model where c is fixed would be worth exploring.

4.5.1 The 2-Parameter Expectancy Valence Model

Assuming consistent effort across trials implies that the Consistency parameter c is no longer required as a measure of boredom or fatigue. However, the choice of value for c will have an impact on the sensitivity function displayed in Equation 4.5 which can not be ignored. Choosing $c = 0$ will render $\theta(t) = 1$, removing the impact of the sensitivity function all together, while $c = 1$ produces $\theta(t) = t/10$ resulting in linearly increasing sensitivity to Expectancies across trials. Neither increasing sensitivity linearly across trials or removing sensitivity to Expectancies all together are realistic reflections of how participants are expected to respond to outcomes on the IGT. A value of $0 < c < 1$, however, results in a dampening effect where the rate of increase in θ is relatively rapid for early trials, but decreases as the trials continue. Incorporating this into the probability of choice (Equation 4.4),

choosing a value of $0 < c < 1$ models a lack of confidence in the Expectancies for very early trials, but it also dampens the effect of the sensitivity function $\theta(t)$ as t becomes larger. Although enforcing an early lack of confidence coupled with long-term dampening can be achieved, to different degrees, for all $0 < c < 1$, $c = 0.5$ also has some empirical evidence for its use. For three of the four groups presented in Experiment 1, $c = 0.5$ was within the 95% credible intervals for the predicted value of c (Table 4.2). In addition, when estimating across all of the participants irrespective of group, the 95% credible interval for the predicted value of c across all of the data also contained $c = 0.5$. This would suggest that, not only is $c = 0.5$ justifiable theoretically, but it is also supported by the data presented in this study.

Due to theoretical and empirical evidence supporting its use, the following analysis was completed using a 2-parameter EVM consisting of the sensitivity function;

$$\theta_{new}(t) = \sqrt{\frac{t}{10}} \tag{4.8}$$

for trial t, along with Equations 4.1,4.2 and 4.4. This is equivalent to the original EVM with the c parameter set to $c = 0.5$.

4.5.2 Method

Participants

The data used for this analysis consists of the same data presented in Experiment 1, Section 4.3.1.

To test the accuracy of the 2-parameter EVM implementation of the model, IGT data was simulated for 361 synthetic participants across every possible combination of 19 systematic values each for W and ϕ for the 2-parameter EVM. The 2-parameter EVM was then fit to the generated data using the formulas described above to see if the imputed parameter values

could be recovered. As with Experiment 1, posterior predictive checks were also completed.

4.5.3 Results

The posterior distributions of the parameter estimates for the EVM for all of the 25 real participants, with trace plots to display the mixing of the chains, are presented in the supplementary material. For the 2-parameter EVM analysis, Raftery's dependence factor, I, was close to 1 for both parameters suggesting that autocorrelation between posterior samples was not a problem and that the chains had mixed well. In addition, all MPSRF values were less than or equal to 1.1, confirming convergence of the chains for every participant.

Parameter Recovery and Posterior Predictive Checks

From the 361 synthetic participants (with known values of W and ϕ) the credible intervals of the posteriors contained the imputed parameter value of $W \simeq 97\%$ and $\phi \simeq 96\%$ of the time. This suggests that the Bayesian estimation of the EVM described here is able to recover the pre-defined parameter values of W and ϕ with a level of accuracy similar to that of the traditional EVM. Posterior predictive checks for the 2-parameter EVM (Figure 4.21) revealed the net return for each participant was always within an expected 95% credible interval for the net return, which is an improvement on the results displayed in Figure 4.6. In addition, comparison of Figure 4.21 to Figure 4.6 shows the mean predicted net return for the participants was closer to the actual net return achieved when using the 2-parameter EVM ($MSE = 221$) than when using the traditional EVM ($MSE = 472$). Despite this, the span of expected net returns covered by the credible intervals had not reduced when compared to Experiment 1 (approximately \$9000, see Figure 4.21). As with Experiment 1, this result suggests that there is a large

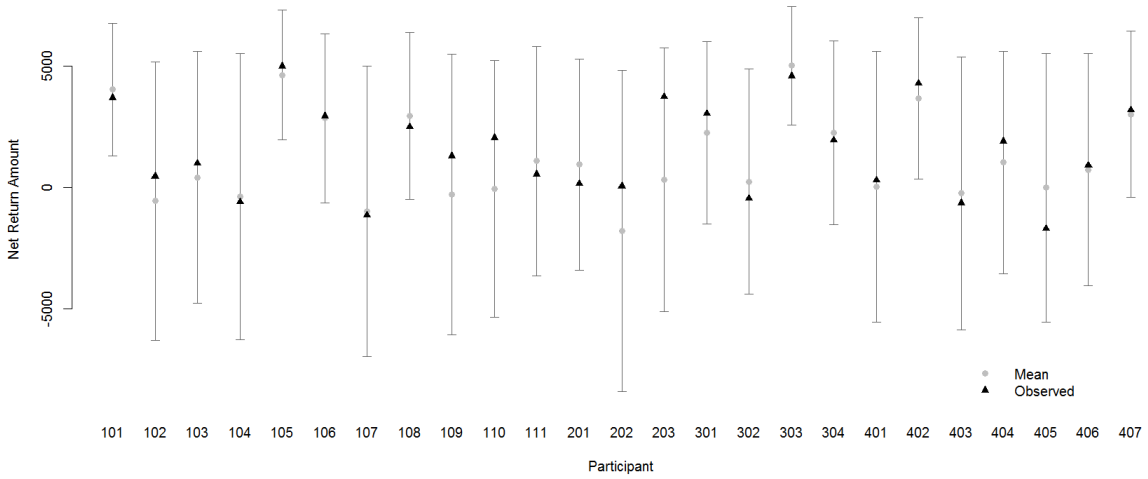

Figure 4.21: 95% Credible intervals for the net returns for each of the 25 participants, generated from posterior predictive checks using the 2-parameter EVM. Here all of the observed net return values were with the predicted 95% Credible intervals for the net returns, but the variability of the credible intervals is still extremely large. As with the results observed in Figure 4.6, this would suggest that there is uncertainty surrounding the predictions from the model.

amount of uncertainty associated with estimates gained using the EVM that may make it impossible to recover behaviour accurately and that reducing the number of parameters does not fully address this problem.

Participant Based Posterior Results and Psychological Inference

Compared to the results using the traditional EVM in Experiment 1, using the 2-parameter EVM resulted in all but 3 of the 25 participants receiving posterior distributions that were reasonably localised and provided concise estimates of W and ϕ. The remaining 3 participants obtained posterior estimates of W with long tails stretching across all possible values of W. However, there was no evidence of bi- or multi-modality, and there was also no observable non-linearity in the joint posteriors for any participant.

Despite the generally clear and concise results, there were 6 out of the 25 participants where the estimates for ϕ obtained using the 2-parameter EVM had a completely contradictory psychological interpretation than those

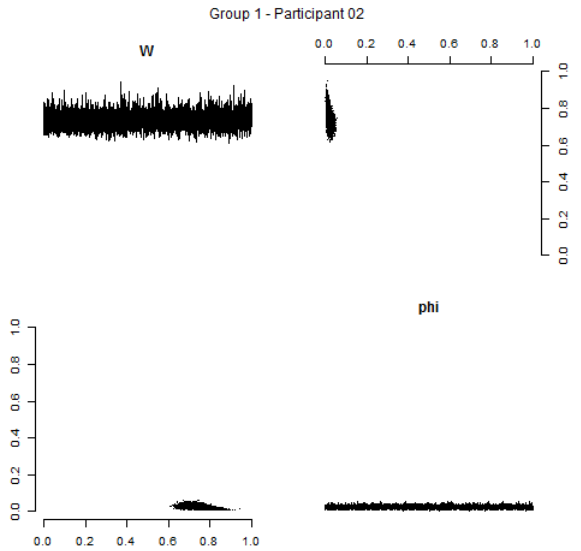

Figure 4.22: The sampled posterior distribution of a drug-naïve participant using the 2-parameter Expectancy Valence Model. This shows estimates of ϕ that are close to zero, which is in direct contrast to the sampled posterior using the traditional EVM (Figure 4.13) in which this participant received estimates of ϕ close to one.

obtained using the traditional EVM. Figure 4.22 shows that using the 2-parameter EVM produced estimates of ϕ close to zero for the second participant of group one, suggesting Expectancies dominated by previous experiences. When using the traditional EVM (Figure 4.13), the same participant received an estimate close to one, suggesting a reliance on the most recent outcomes to inform their Expectancies. The same is true for the other five participants who all received estimates close to zero using the 2-parameter EVM and estimates close to one using the traditional EVM.

In addition, when using the traditional EVM, Figure 4.12 shows a posterior spanning the entire range of ϕ and Figure 4.7 shows a bi-modality in ϕ while both obtained clear estimates of ϕ close to zero when using the 2-parameter EVM (Figures 4.23 and 4.24 respectively). In fact 16 of the 25 participants had credible intervals spanning $0 \leq \phi < 0.1$ which could suggest that the majority of participants formed Expectancies that were not heavily influenced by recent events.

In general, the psychological interpretation of the parameter estimates

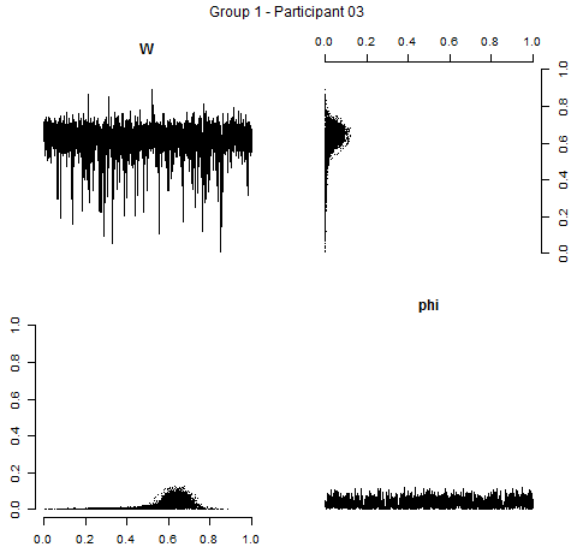

Figure 4.23: The sampled posterior distribution of a drug-naïve participant using the 2-parameter Expectancy Valence Model. This shows estimates of ϕ that are close to zero, compared the sampled posterior using the traditional EVM (Figure 4.12) which shows this participant with an estimate of ϕ spanning the entire range of values.

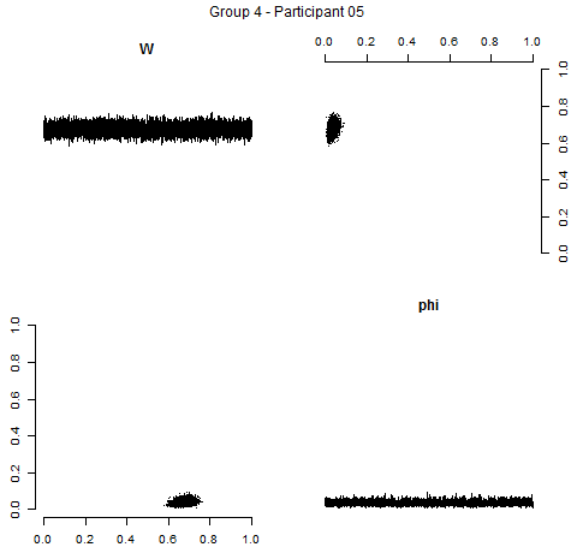

Figure 4.24: The sampled posterior distribution of a participant from the ecstasy-and-cannabis use group using the 2-parameter Expectancy Valence Model. This shows estimates of ϕ that are close to zero, compared the sampled posterior using the traditional EVM (Figure 4.7) which shows this participant with bi-modal estimates of ϕ.

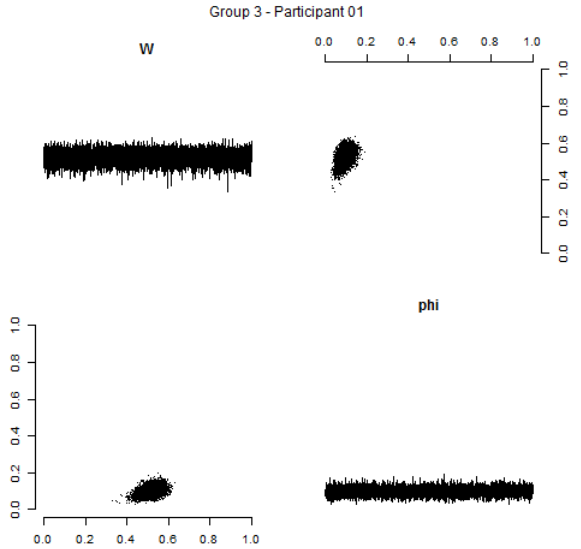

Figure 4.25: The sampled posterior distribution of a participant from the cannabis use group using the 2-parameter Expectancy Valence Model. This shows concise posterior distributions that have a clear psychological interpretation.

derived using the 2-parameter EVM are clearer and tell a more consistent story than the traditional implementation of the EVM. The posterior in Figure 4.22, for example, suggests an individual with reasonable memory but with a tendency to focus more highly on wins than losses, where as Figure 4.25 suggests someone with a more balanced consideration of win and loss experiences. A such, by producing narrower, reasonably localised posteriors, the estimates at the level of the individual are more exact and easier to interpret.

Group Estimates

In general, there was more consistency between participants within all of the groups than the results displayed in Experiment 1. Despite the increased consistency within groups, Figures 4.26, 4.27 and 4.28 show there were still some individuals with estimates of W lying outside of the group estimates displayed in Table 4.3. These relative outliers may have contributed to the increased uncertainty observed in the W parameter estimates in Table 4.3 in comparison to the results in Experiment 1 (Table 4.2).

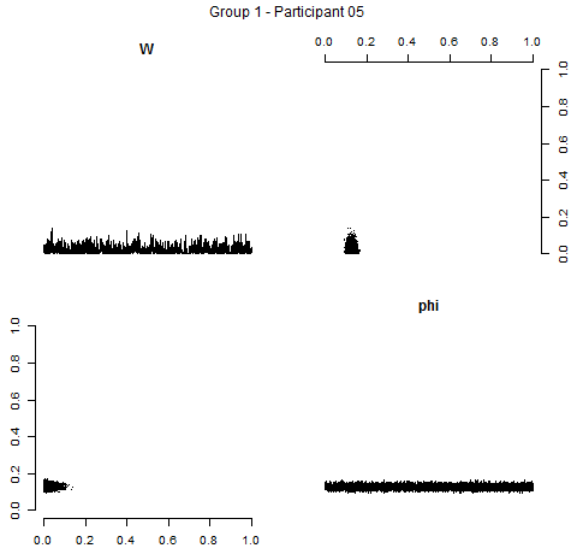

Figure 4.26: The sampled posterior distribution of a drug-naïve participant using the 2-parameter Expectancy Valence Model. The estimate for W for this participant is outside of the estimate for the drug-naïve group displayed in Table 4.3.

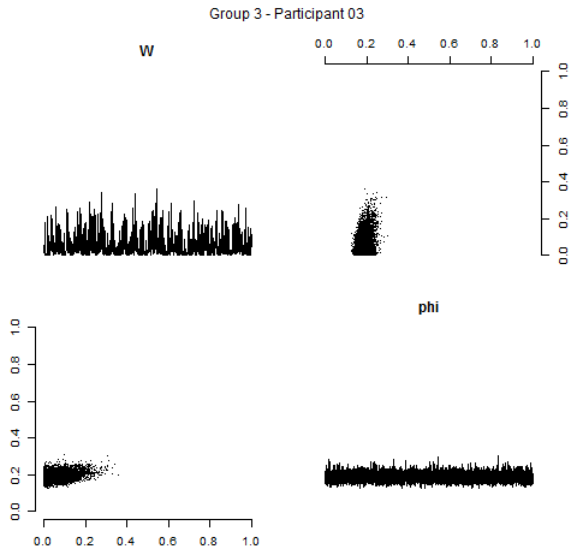

Figure 4.27: The sampled posterior distribution of a participant from the cannabis use group using the 2-parameter Expectancy Valence Model. The estimate for W for this participant is outside of the estimate for the cannabis using group displayed in Table 4.3.

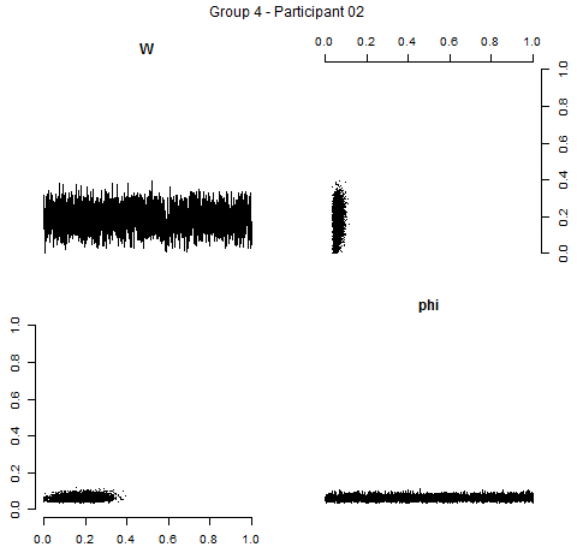

Figure 4.28: The sampled posterior distribution of a participant from the ecstasy-and-cannabis use group using the 2-parameter Expectancy Valence Model. The estimate for W for this participant is outside of the estimate for the ecstasy-and-cannabis use group displayed in Table 4.3.

Table 4.3: Credible intervals for W and ϕ for each of the four drug using groups presented in Experiment 1 and Experiment 3.

	95% credible Interval							
	Experiment 1				Experiment 3			
	W		ϕ		W		ϕ	
Group	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Drug Naïve	0.56	0.77	0.00	0.39	0.33	0.73	0.02	0.09
MDMA	0.24	0.69	0.10	0.69	0.38	0.96	0.00	0.11
Cannabis	0.36	0.73	0.00	0.94	0.10	0.69	0.03	0.24
MDMA & Cannabis	0.36	0.70	0.00	0.32	0.30	0.77	0.01	0.08

4.5.4 Summary

Using a 2-parameter EVM to interpret performance on the IGT resulted in more localized, uni-modal, approximately normal estimates of parameter distributions than using the traditional EVM. These results were more clearly psychologically interpretable than those presented in Experiment 1 as there was no evidence of bi-modality or non-linearity in the posteriors. As such, the 2-parameter EVM is able to provide clearer, more easily interpretable findings that the traditional EVM when used as to interpret performance on the IGT.

However, part of the problem associated with the traditional implementation of the EVM was that it was not always able to distinguish between different explanations of behaviour. Specifically, it could not distinguish between a bored participant with great memory or a participant who was trying really hard but had terrible memory. In the 2-parameter implementation of the EVM, results were clearer and more easily interpretable but, due to the reduced number of parameters, it is still unable to distinguish between these different explanations of behaviour.

Posterior predictive checks of the implementation of the 2-parameter EVM also suggested that there may be a lack of consistency in net outcomes when using the IGT. This is in line with the findings from Experiment 1 which also showed highly variable net outcomes were possible when using the same parameter values to simulate performance on the IGT. The reliability of the IGT is, therefore, something that would benefit from further investigation.

4.6 Discussion

The presented results highlight two main issues, the first of which is with the reliability of the IGT itself. Both Buelow and Suhr (2009) and Lin et al. (2013) raised concerns about the reliability of the IGT and, with the lack of consistency in net return experienced on the IGT across the three Experiments presented here, the current study agrees that the reliability of the IGT needs further investigation. In addition, due to the observed lack of consistency on repeated measurement within the same individual, we do not recommend using frequency of deck choice or net outcomes to measure individual performance on the IGT.

If basic analysis of performance on the IGT, such as net return and frequency of deck choice, are not recommended, then cognitive models of

performance such as the EVM must be employed. This leads to our second main issue: That the EVM does not clearly measure the underlying psychological processes driving observed performance on the IGT.

Irrespective of whether a participant completed the IGT once or multiple times, the EVM, when used to decompose performance on the IGT, did not produce the localized, uni-modal or approximately normal parameter distributions required to accurately summarise behaviour. The distributions produced were often so variable as to include every possible parameter value, rendering psychological interpretations of posterior distributions inconsistent and contradictory. These multiple, conflicting parameter estimates suggest that the EVM is not always able to distinguish between different explanations of behaviour. In addition, conflicting outcomes between members of the same group were observed, along with summaries of group performance that contradicted the individual outcomes of the members who made up that group. Both of these results are further evidence that the EVM is not always able to distinguish between different explanations of behaviour.

When reducing the traditional EVM to a 2-parameter EVM, in which the consistency parameter is set at $c = 0.5$, improvements in the uncertainty of estimates were observed. The 2-parameter EVM produced more localized, uni-modal, approximately normally distributed estimates of parameter distributions than using the traditional EVM, allowing for clearer psychological interpretation of results. There was also greater consistency between results for participants in the same psychological group when using the 2-parameter EVM, leading to group estimates that represented the diversity of behaviours present in the group more fully than when using the traditional EVM. Despite these improvements, there was still some uncertainty surrounding estimates of the W parameter when using the 2-parameter EVM

and further investigation into reduced parameter models would be advisable.

Highly variable individual-level parameter estimates could be evidence of mathematical models that do not measure the behaviour they are assumed to measure. The 2-parameter EVM presented here has gone some way towards investigating if this is the case with the EVM when applied to IGT data. However, Chiu and Lin (2007) suggest that the IGT may not be measuring the behavioural deficits it was designed to measure and, if this is the case, then the EVM cannot be expected to reliably capture those traits. Apart from the issues identified in this chapter in relation to the reliability of the IGT, Steingroever et al. (2013a) suggest that there may also be a flaw in the fundamental assumptions associated with the creation of the IGT. Steingroever et al. (2013a) show that even healthy participants display highly idiosyncratic choice behaviours when completing the IGT, challenging the assumption by Bechara et al. (1994) that healthy participants initially explore the decks and then move to the behaviour of exploiting the decks they have learned are the best. Steingroever et al. (2013a) suggest that individuals do not move away from the initial exploration stage when completing the IGT and hence highly idiosyncratic choice behaviours, like those observed in the current study, predominate (Figures 4.2 and 4.5). As such, it would appear that the decision making processes required to complete the IGT successfully are more complicated than first thought. Even recent alternatives to the EVM such as the Prospect Valence Model (Ahn et al., 2008, 2011) have been unable to provide good fits to all patterns of choice behaviours (Steingroever et al., 2013b).

The findings presented in this and recent studies show clear evidence that there may be problems with both the EVM and the IGT and, to improve the clinical utility of this popular task, it is worth investigating how to improve them both. The current study provides evidence that highlights a potential

lack of reliability in the IGT itself. In addition, using simple measures of performance on the IGT, such as net return or frequency of deck selection, was found to be an ineffective way to measure decision making deficits in individuals or at group level. The EVM, when used to measure deficits in the underlying psychological processes required for completion of the IGT, provides multiple, conflicting parameter estimates which suggest that the EVM is not always able to distinguish between different underlying reasons for observed behaviour. As such, the current study warns against its use at both the level of the individual and at the level of the group. A reduced parameter EVM, such as the 2-parameter EVM presented in this chapter, may provide clearer and more consistent results than using the EVM. However, a reduced parameter EVM can not address the inability of the EVM to distinguish between different, competing psychological explanations of behaviour and investigations into a model that may be able to address this issue would be supported.

Chapter 5

Supplementary Material for *The Expectancy Valence Model of the Iowa Gambling Task: Can it produce reliable estimates for individuals?*

The Supplementary Material provides the deck selection profiles and posterior distributions (with mixing) for all three experiments and posterior predictive checks and group posterior distributions for Experiment 3 are provided.

5.1 Posterior Predictive Checks

Posterior predictive checks test whether a model is an adequate description of the observed data it is trying to describe. The steps of a Posterior predictive check can be summarised as follows:

1. Fit the model to the observed data and gain posterior distributions for the model parameters.

2. Draw samples from the posterior distributions for the parameters.

3. Using the samples from the posterior, generate (using the model) simulated data sets.

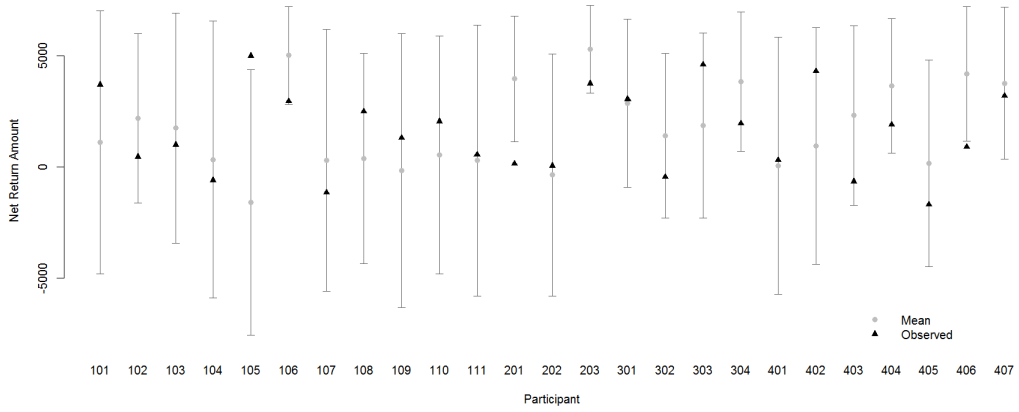

Figure 5.1: Posterior predictive check results the EVM as applied in Section 4.3.4.

4. Compare the simulated and observed data sets on a reasonable test statistic that summarises the data.

In this way it is possible to check if the model describes the data well.

In this chapter, posterior predictive checks were completed by randomly sampling $1,000$ parameter values from the posteriors gained for the parameters of the EVM and 2-parameter EVM for each of the 25 participants. The sampled parameter values were then used to simulate $1,000$ data sets that mimicked a 250 choice IGT trial. Let's call each of the 25 sets of $1,000$ IGT runs (one for each participant) the "simulated data". The mean and 95% confidence interval was then obtained for the net return for each of the simulated data and this was compared to the net return obtained by the true participant.

5.2 Experiment 1

5.2.1 Posterior Predictive Checks

The results are displayed in Figure 5.1:

Figure 5.1 shows that, for all but three of the participants, the observed net return is within the confidence interval for the net return generated from the simulated data. This would imply that the model fits the data

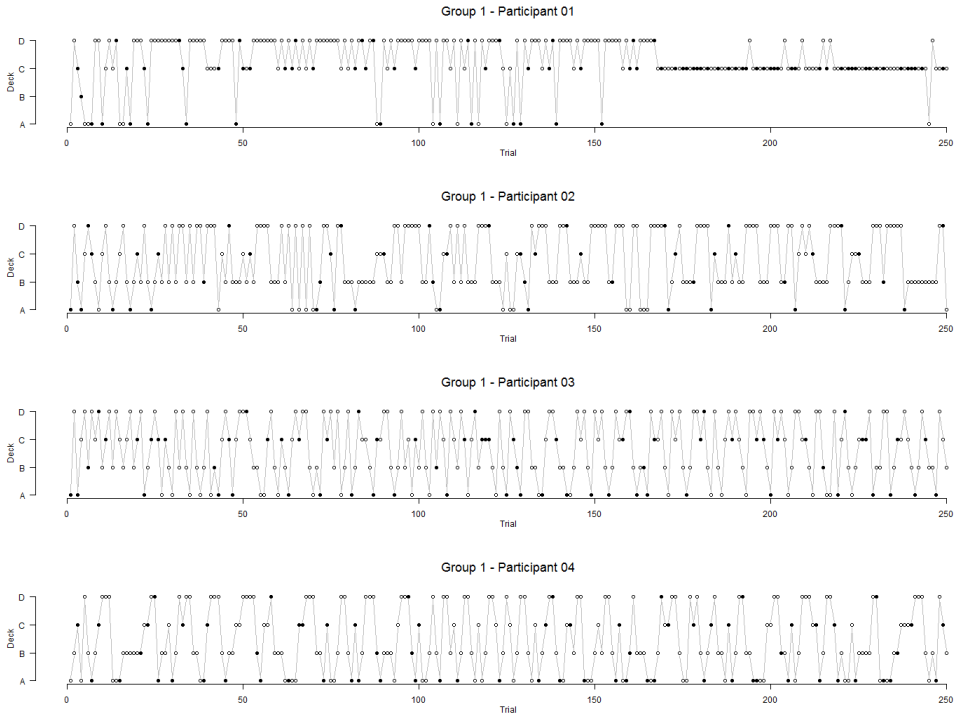

Figure 5.2: Deck selection profiles for Participants 101 through 104.

well, however looking at the values on the y-axis is cause for concern. On average, the confidence intervals for the net return from the simulated data span just under $9,000. This would imply that there is a large amount of uncertainty attached to the observed parameters, which is in line with the rest of the results presented in this chapter.

5.2.2 Deck Selection Profiles

Figures 5.2 through 5.8 are the deck selection profiles for all participants in Experiment 1 (Section 4.3). A circle represents a choice from the corresponding deck label on the left of the figure and if the circle is filled, then a loss was experienced. The participant numbers are provided in the header for each figure. The first number of each participant number indicates group inclusion where group 1 are drug naïve controls, group 2 are regular users of MDMA, group 3 are regular users of cannabis and group 4 are regular users of both MDMA and cannabis. Each participant completed the IGT once with 250 deck selections.

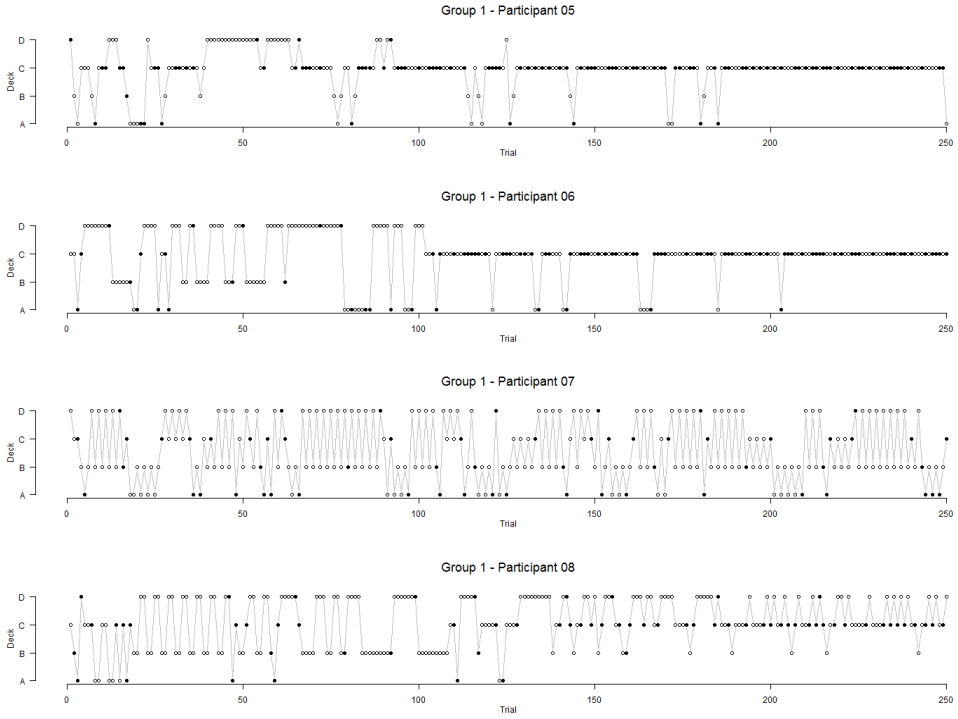

Figure 5.3: Deck selection profiles for Participants 105 through 108.

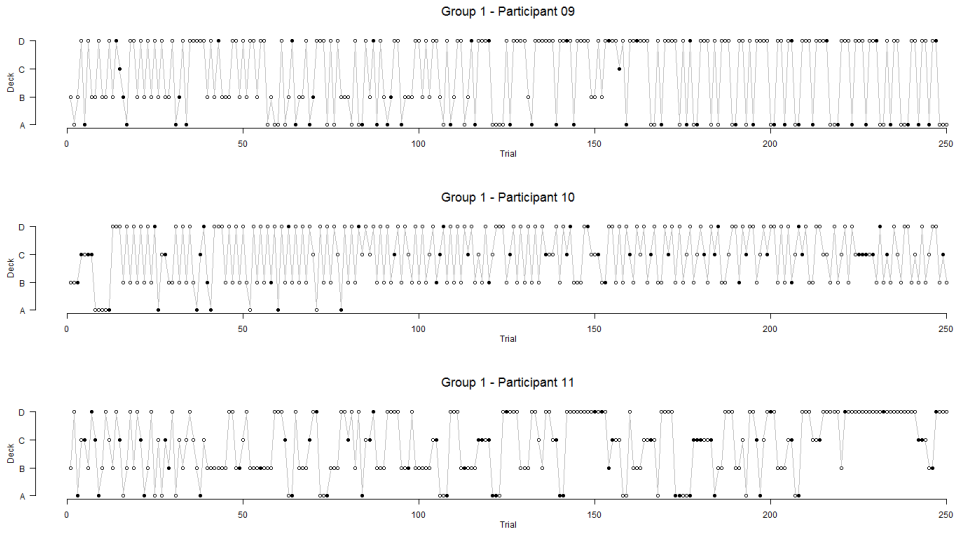

Figure 5.4: Deck selection profiles for Participants 109 through 111

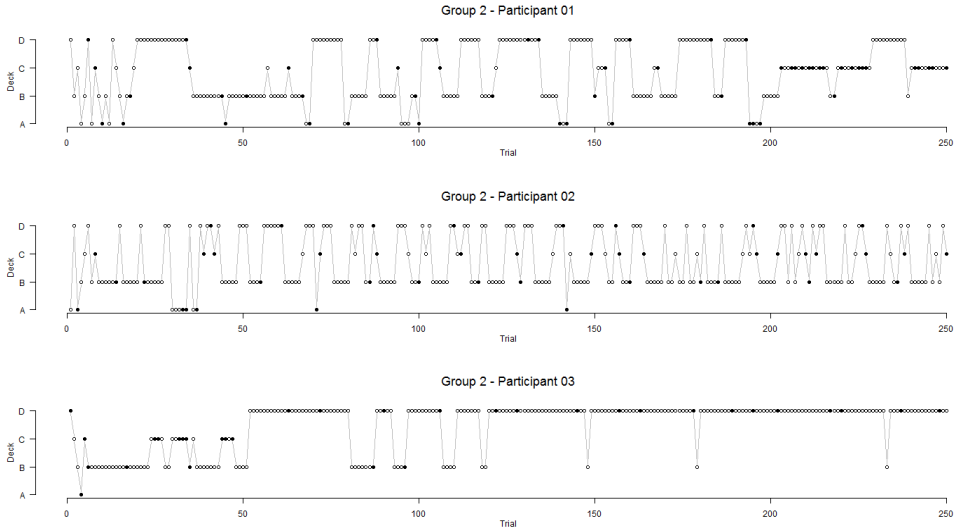

Figure 5.5: Deck selection profiles for Participants 201 through 203.

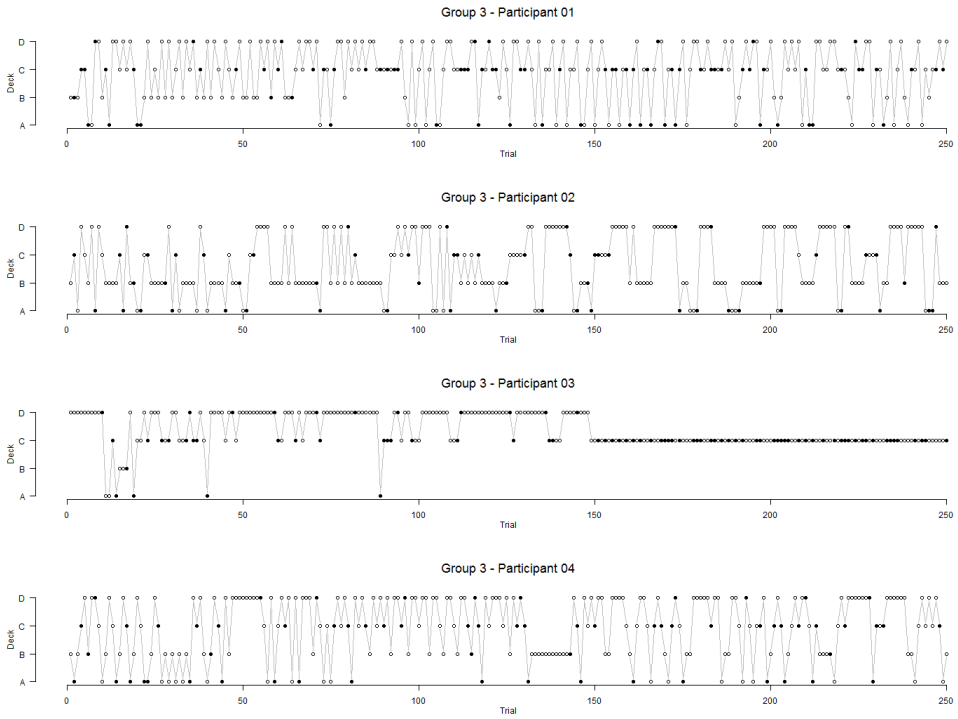

Figure 5.6: Deck selection profiles for Participants 301 through 304.

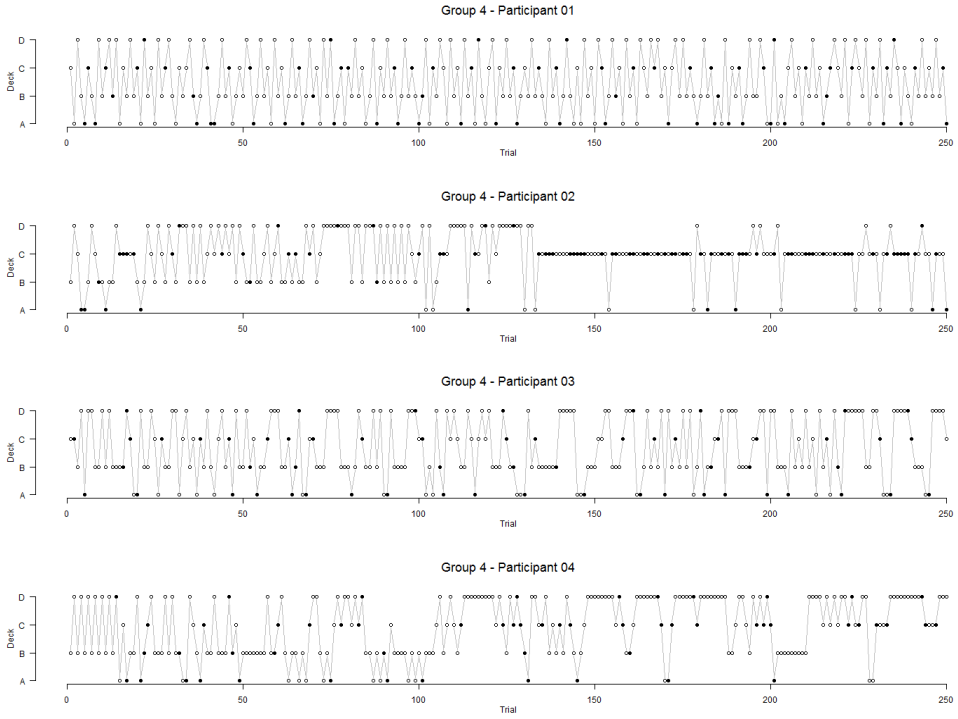

Figure 5.7: Deck selection profiles for Participants 401 through 404.

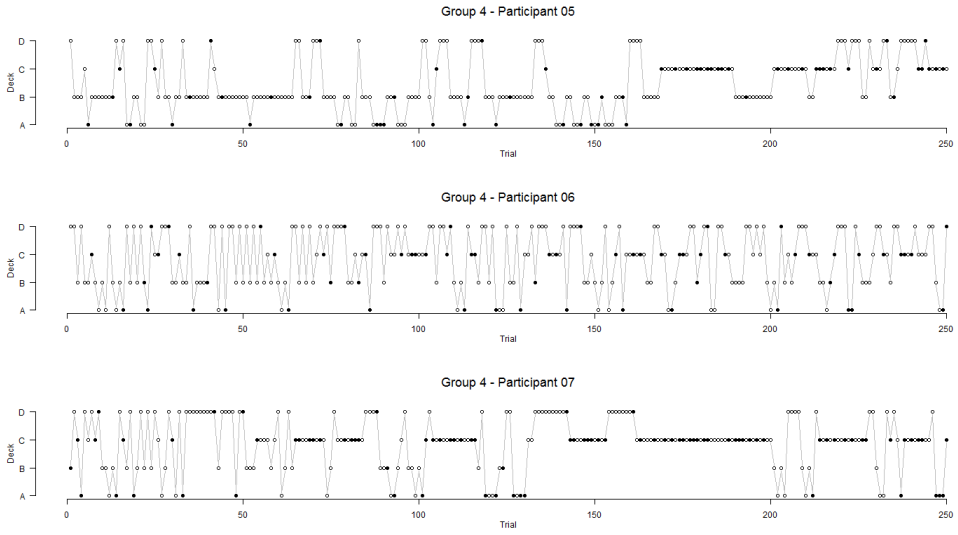

Figure 5.8: Deck selection profiles for Participants 405 through 407.

5.2.3 Posterior Distributions

Figures 5.9 through 5.15 are the sampled posterior distributions for the parameters of the EVM for the all participants completing the IGT in Experiment 1 (Section 4.3). These are pairwise plots in which the three upper diagonal panels are a reflection of the three lower diagonal panels. Across the central diagonal, the mixing for each of the three parameters (W, ϕ and c) is presented. This is the mixing for all five MCMC chains concatenated together such that the first 2,000 samples correspond to the first chain, the second 2,000 samples to the second chain and so on. The participant numbers are provided in the header for each figure. The first number of each participant number indicates group inclusion where group 1 are drug naïve controls, group 2 are regular users of MDMA, group 3 are regular users of cannabis and group 4 are regular users of both MDMA and cannabis. Each participant completed the IGT once with 250 deck selections. In this Experiment, participants completed the IGT once with 250 deck selections.

5.2.4 Group Performance

Figure 5.16 shows the sampled posterior distributions for the parameters of the EVM for each of the four groups completing the IGT in Experiment 1 (Section 4.3). These are pairwise plots in which the three upper diagonal panels are a reflection of the three lower diagonal panels. Across the central diagonal, the mixing for each of the three parameters (W, ϕ and c) is presented. This is the mixing for all five MCMC chains concatenated together such that the first 2,000 samples correspond to the first chain, the second 2,000 samples to the second chain and so on. The group numbers are provided in the header for each figure where group 1 are drug naïve controls, group 2 are regular users of MDMA, group 3 are regular users of cannabis and group 4 are regular users of both MDMA and cannabis.

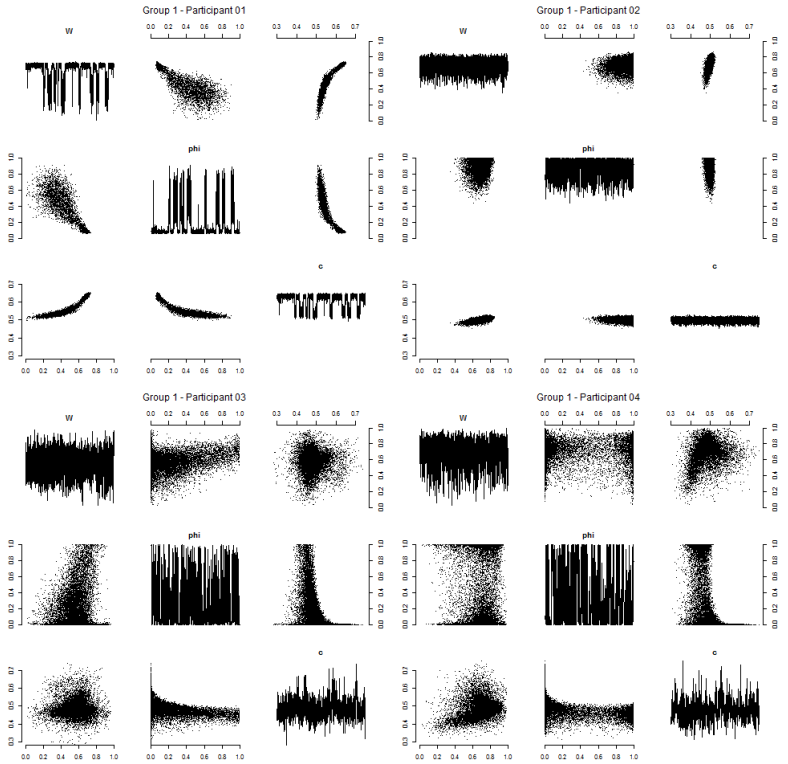

Figure 5.9: Sampled posterior distributions (with mixing) for Participants 101 through 104

.

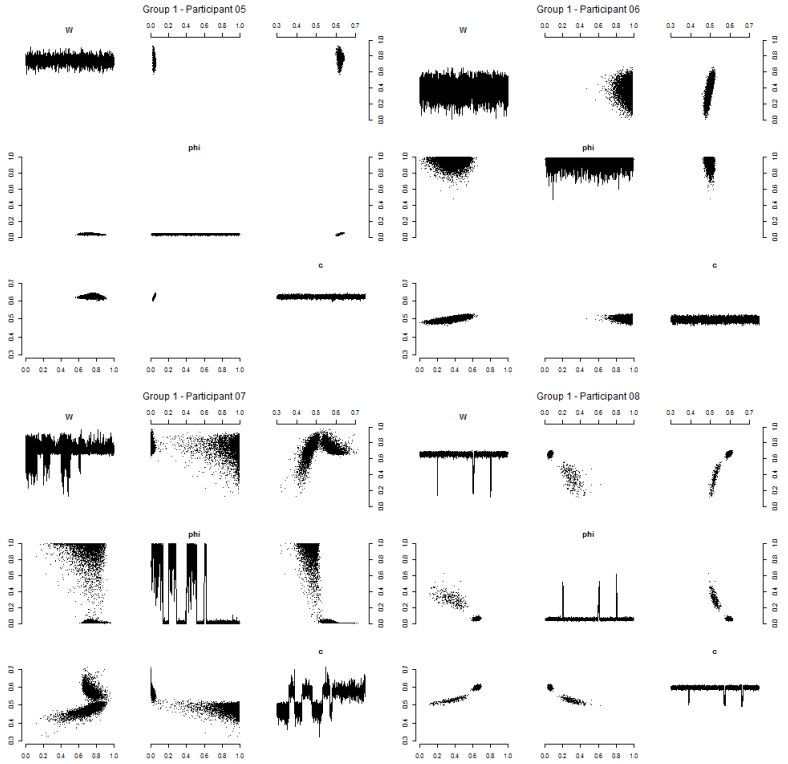

Figure 5.10: Sampled posterior distributions (with mixing) for Participants 105 through 108

.

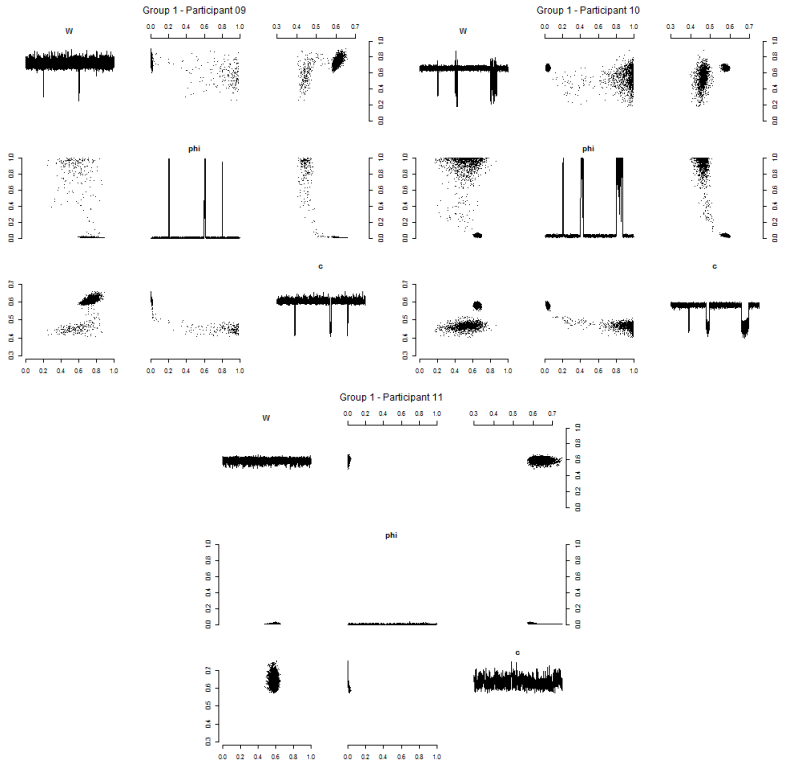

Figure 5.11: Sampled posterior distributions (with mixing) for Participants 109 through 111

.

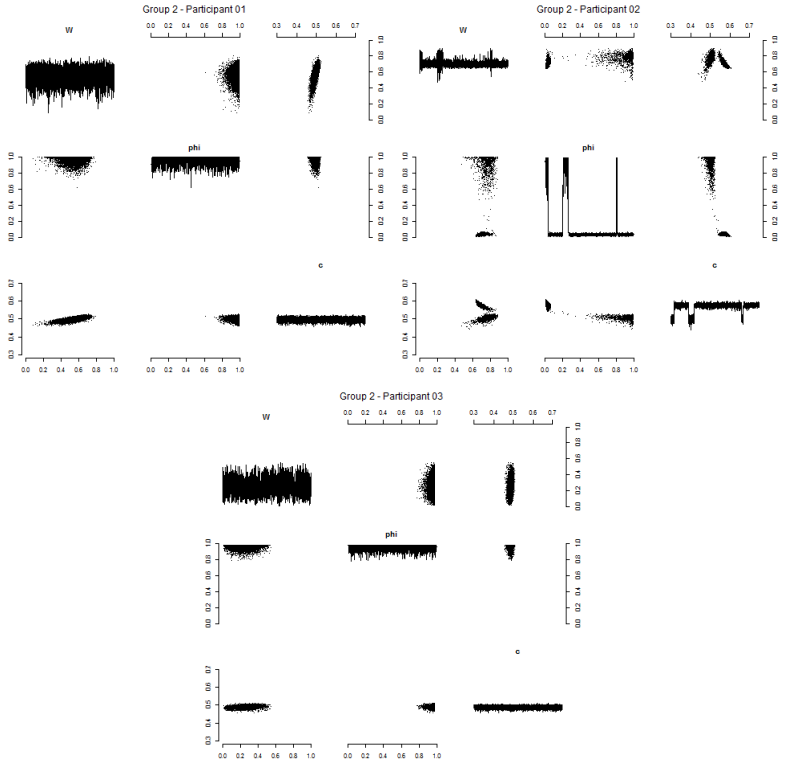

Figure 5.12: Sampled posterior distributions (with mixing) for Participants 201 through 203

.

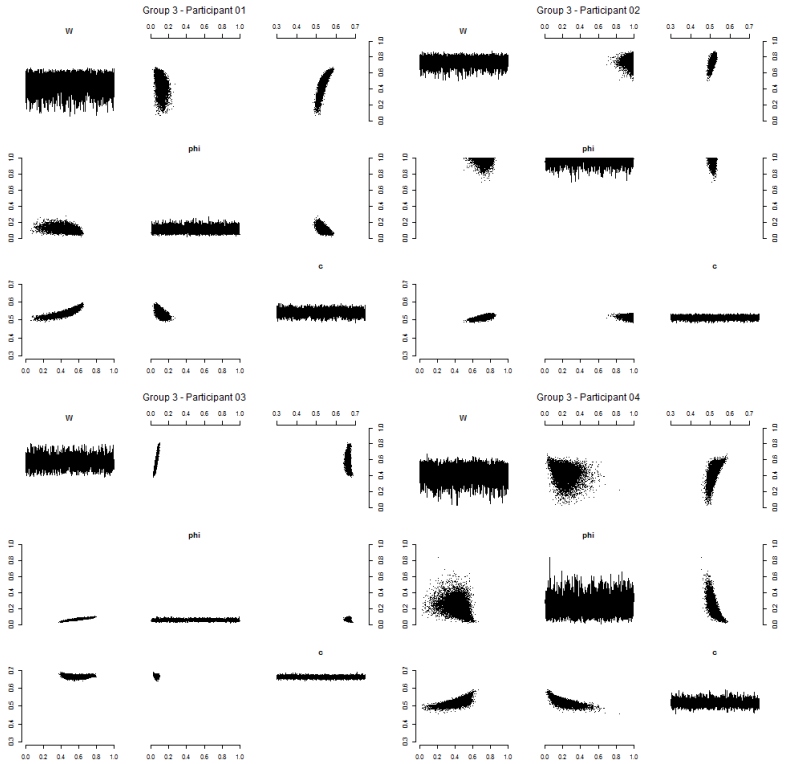

Figure 5.13: Sampled posterior distributions (with mixing) for Participants 301 through 304

.

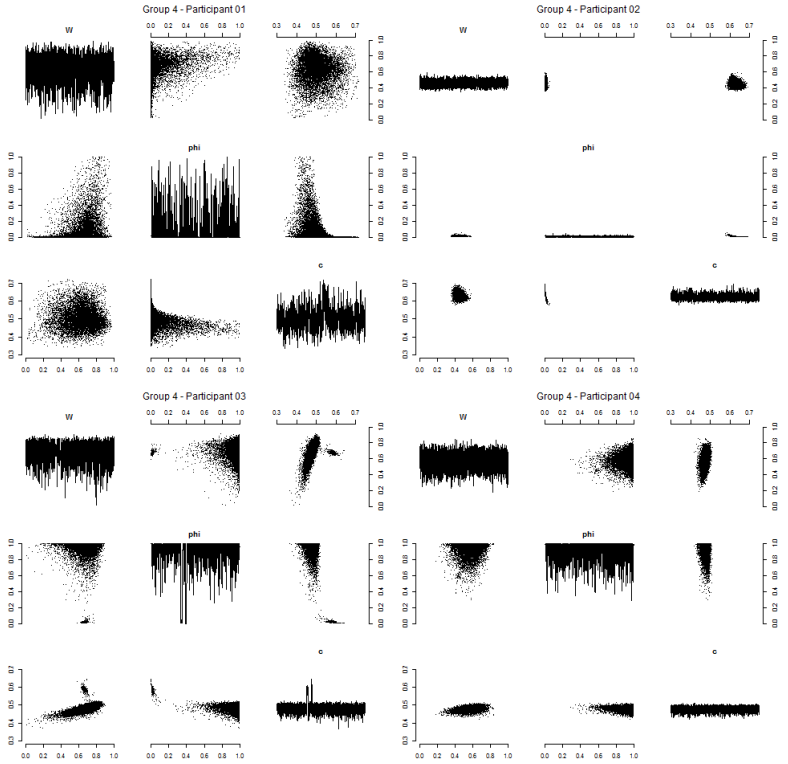

Figure 5.14: Sampled posterior distributions (with mixing) for Participants 401 through 404

.

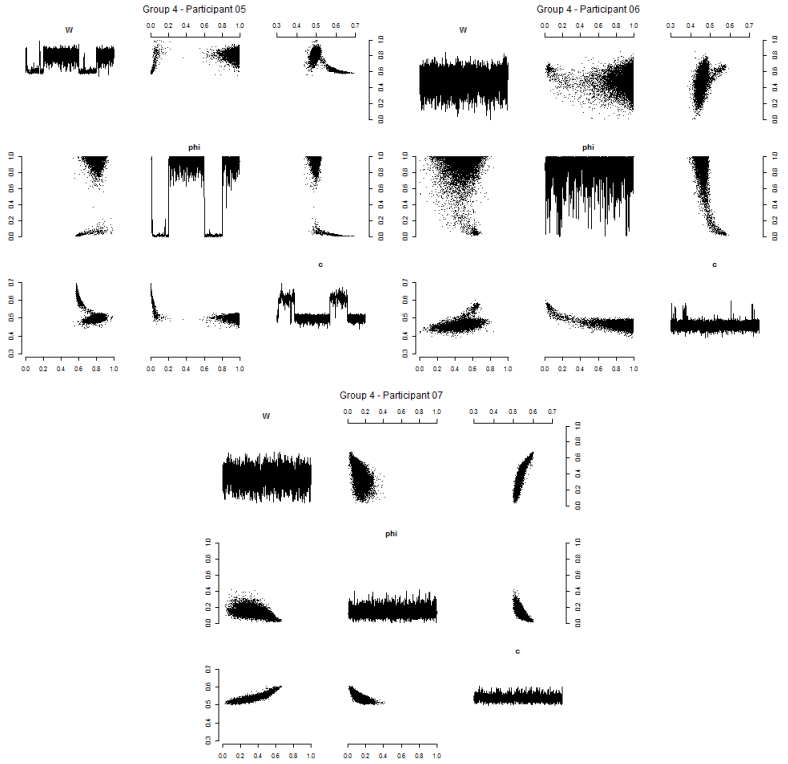

Figure 5.15: Sampled posterior distributions (with mixing) for Participants 405 through 407

.

5.3 Experiment 2

5.3.1 Deck Selection Profiles

Figures 5.17 through 5.24 are the deck selection profiles for all participants in Experiment 2. A circle represents a choice from the corresponding deck label on the left of the figure and if the circle is filled, then a loss was experienced. The participant numbers are provided in the header for each figure. All participants in this study were drug naïve controls. Each participant completed the IGT three times with each of the three runs comprising 100 deck selections. The three runs are each presented separately here.

5.3.2 Posterior Distributions

Figures 5.25 and 5.26 are the sampled posterior distributions for the parameters of the EVM for the all participants completing the IGT in Experiment 2. These are pairwise plots in which the three upper diagonal panels are a

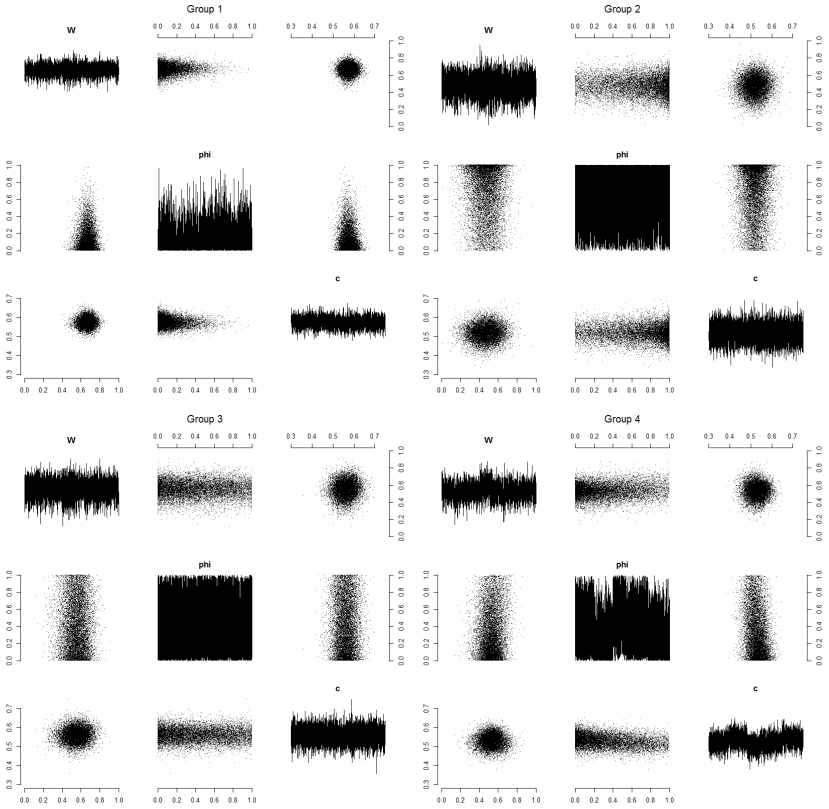

Figure 5.16: Sampled posterior distributions (with mixing) for Groups 1 through 4

.

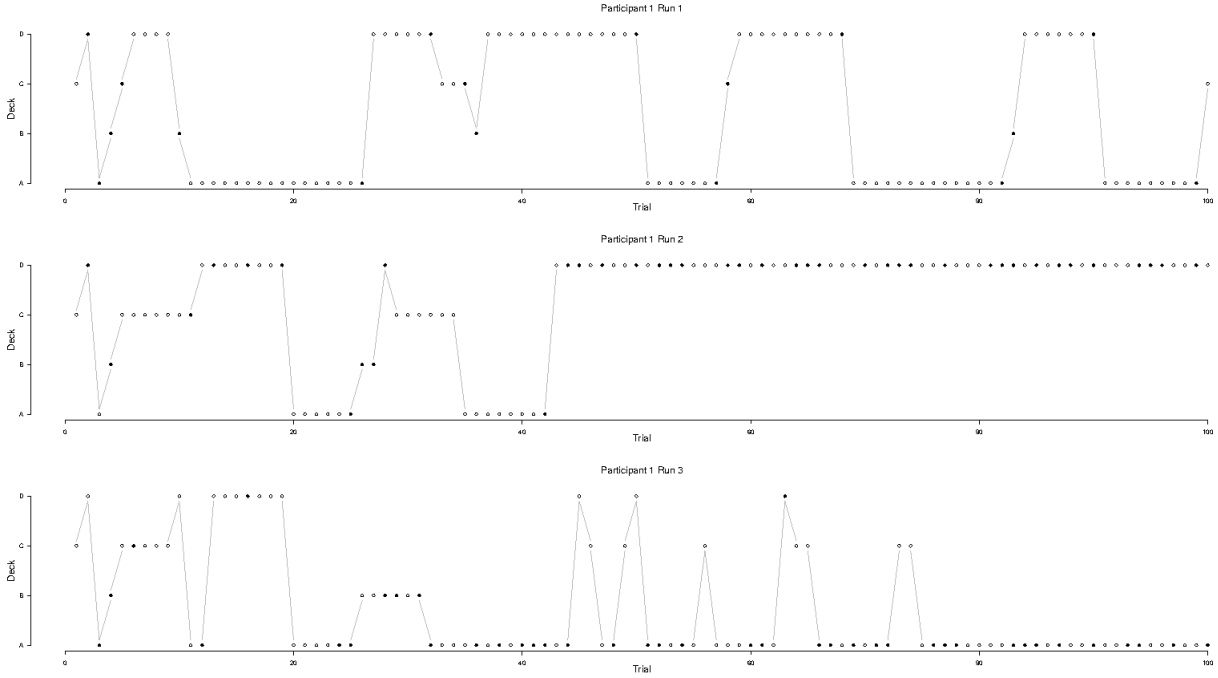

Figure 5.17: Deck selection profiles for all three runs completed by Participant 01.

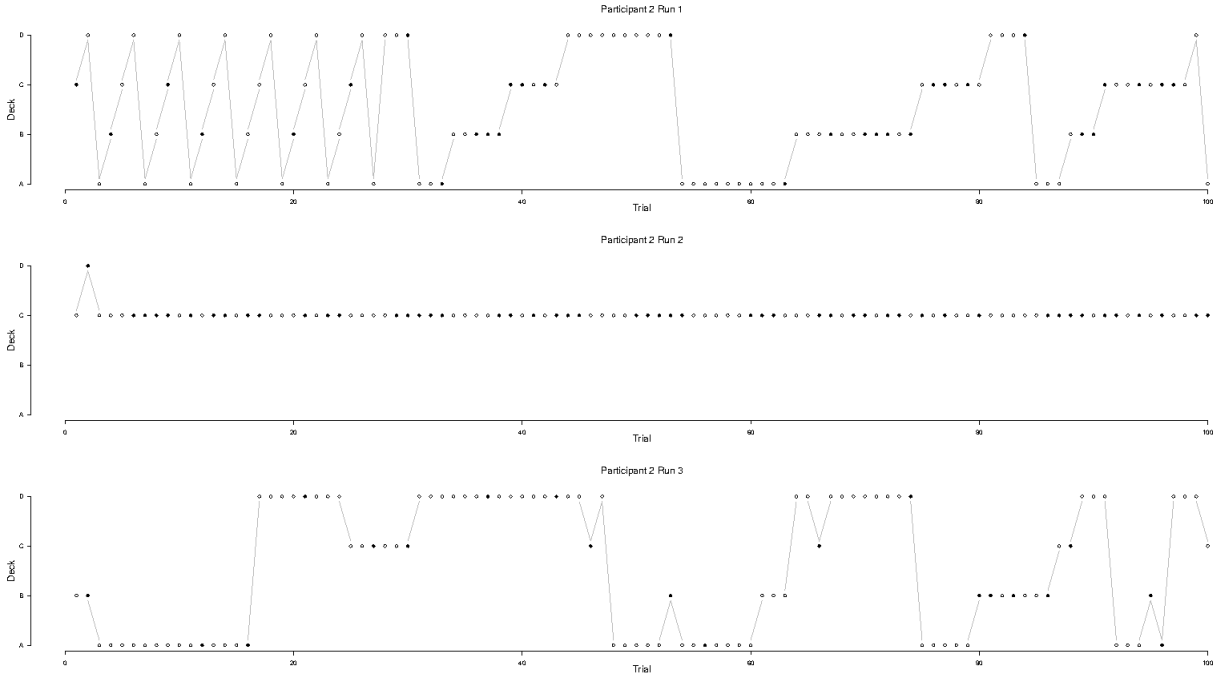

Figure 5.18: Deck selection profiles for all three runs completed by Participant 02.

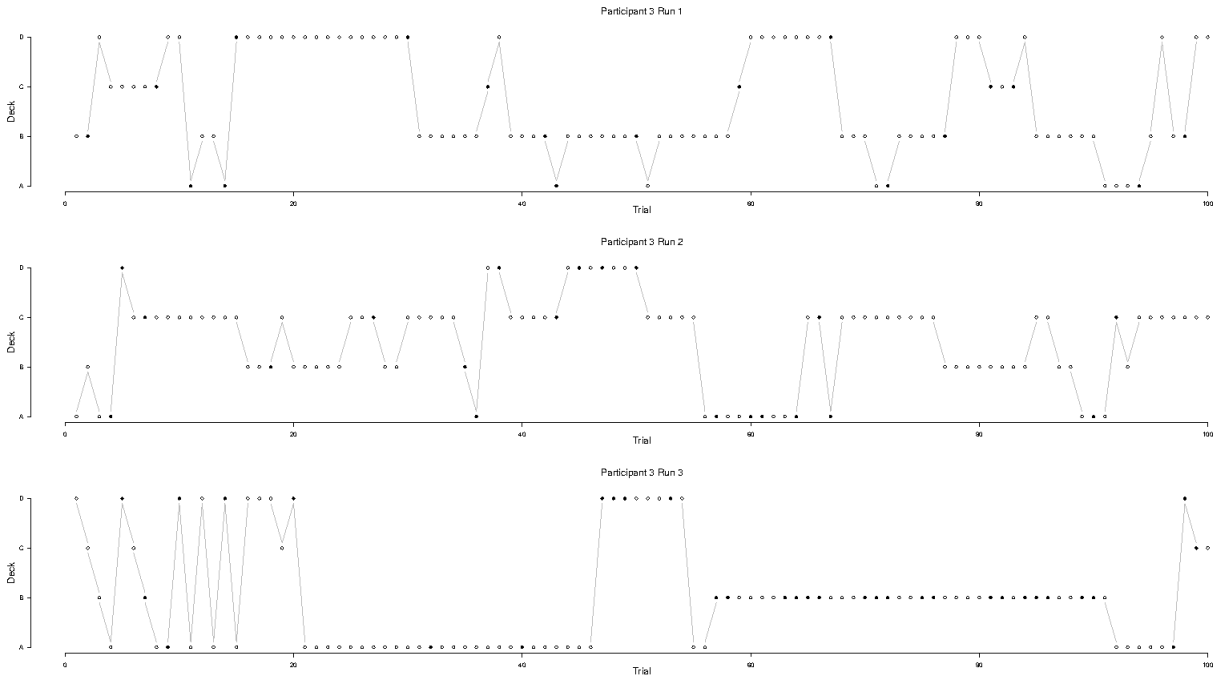

Figure 5.19: Deck selection profiles for all three runs completed by Participant 03.

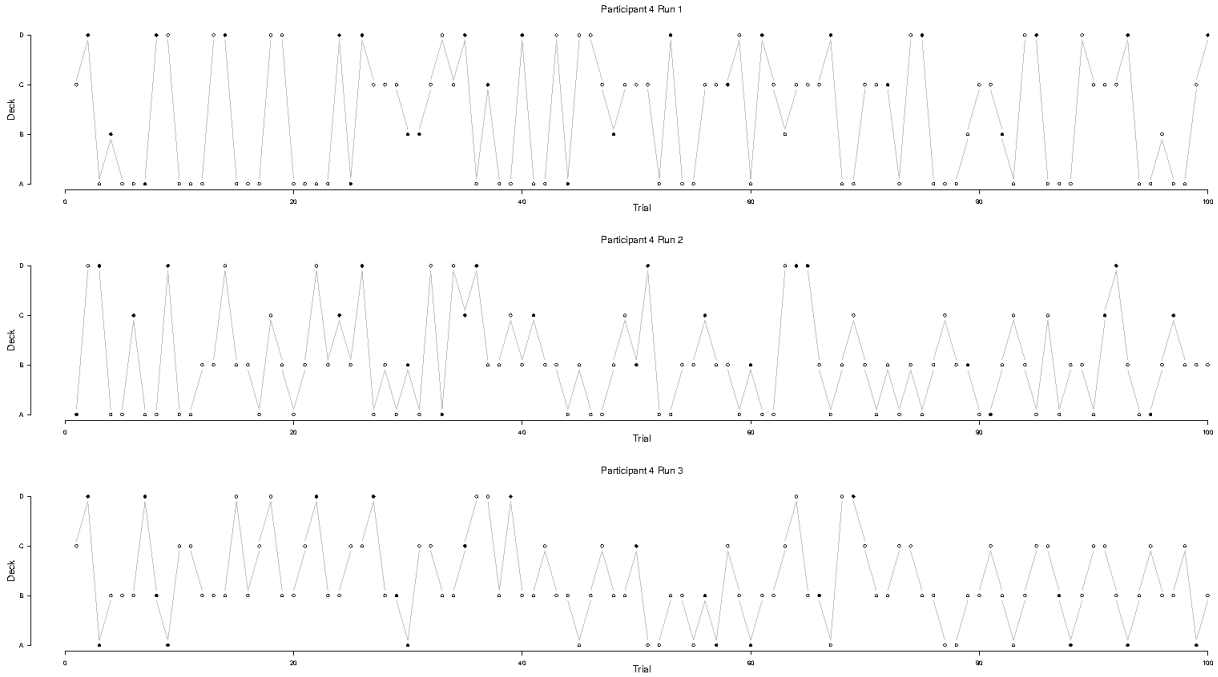

Figure 5.20: Deck selection profiles for all three runs completed by Participant 04.

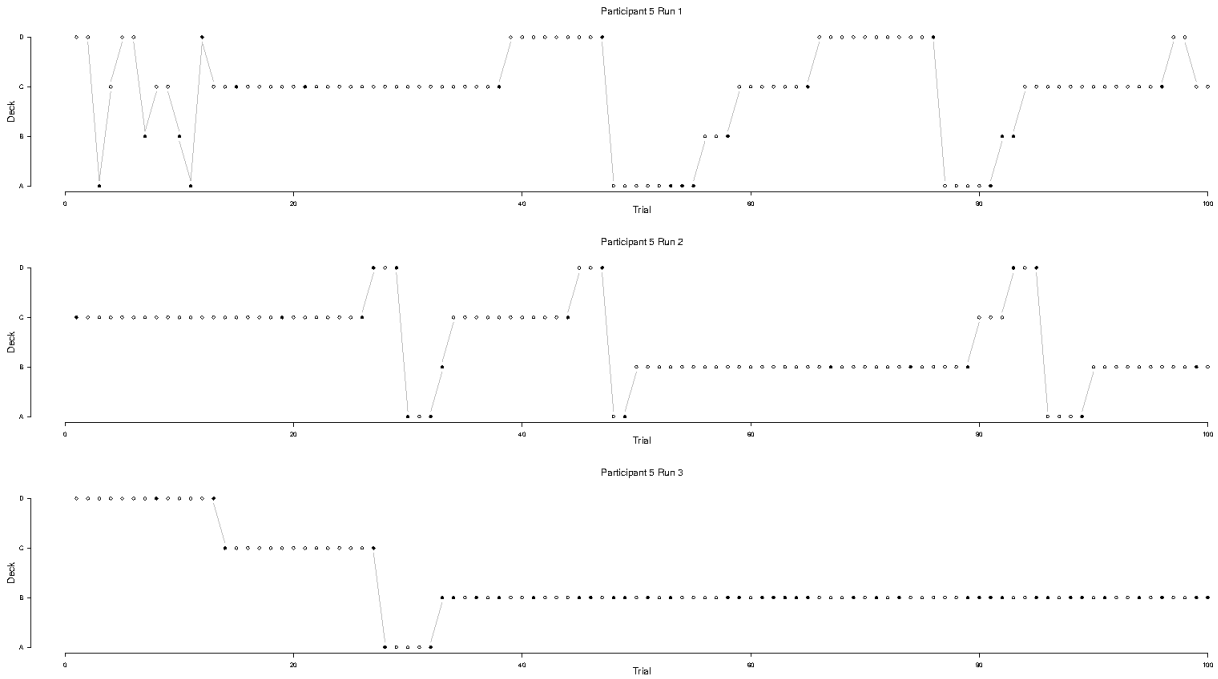

Figure 5.21: Deck selection profiles for all three runs completed by Participant 05.

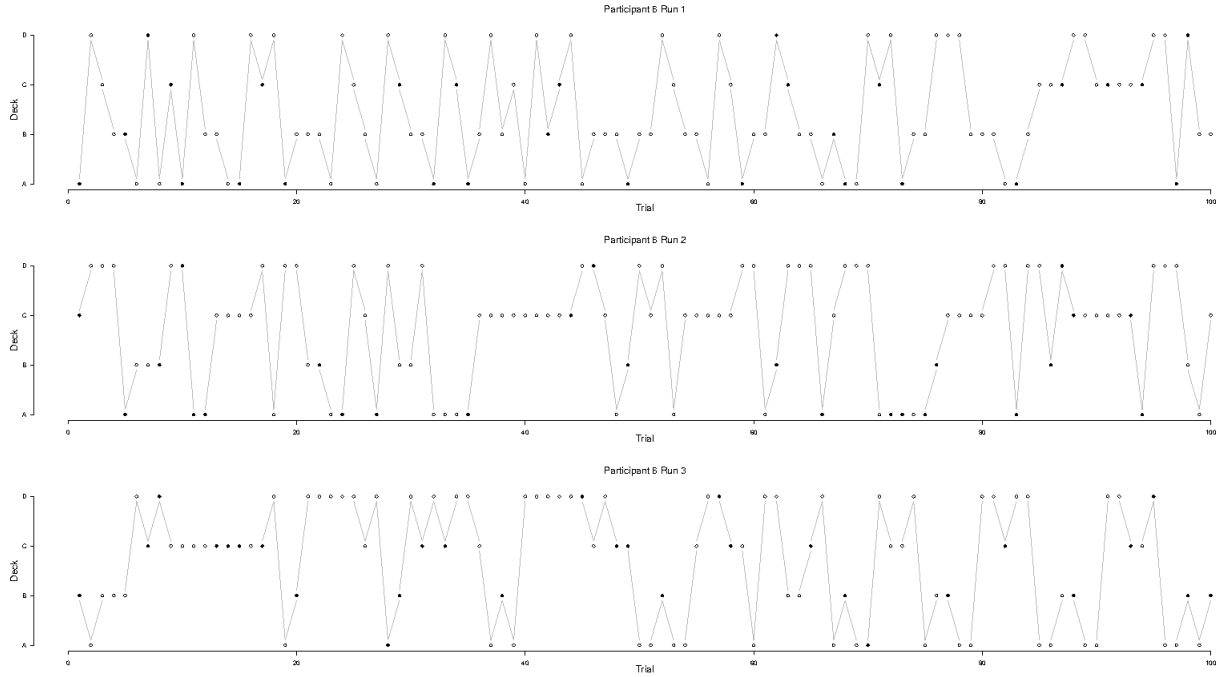

Figure 5.22: Deck selection profiles for all three runs completed by Participant 06.

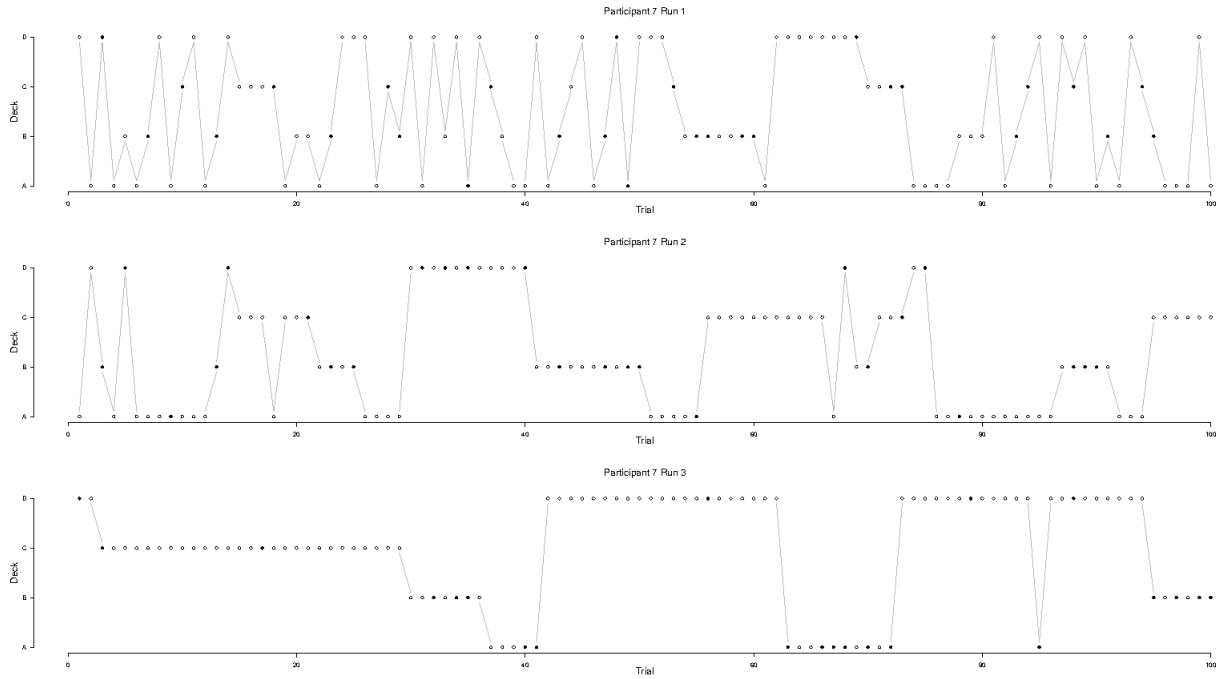

Figure 5.23: Deck selection profiles for all three runs completed by Participant 07.

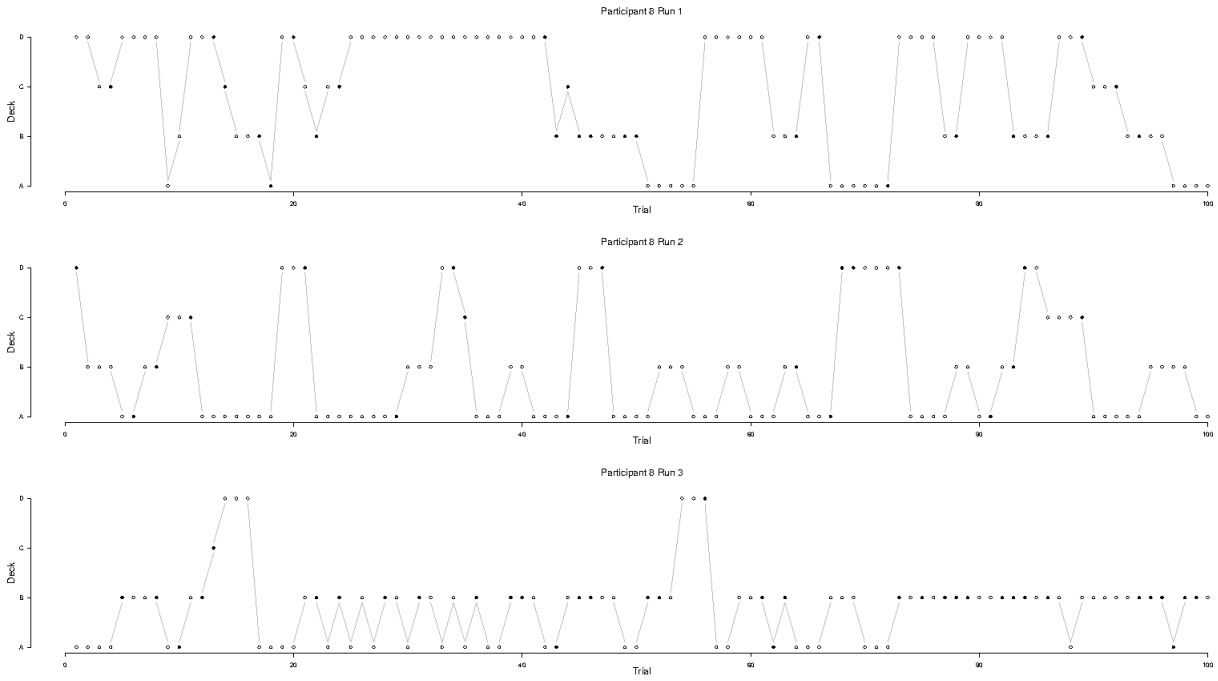

Figure 5.24: Deck selection profiles for all three runs completed by Participant 08.

reflection of the three lower diagonal panels. Across the central diagonal, the mixing for each of the three parameters (W, ϕ and c) is presented. This is the mixing for all five MCMC chains concatenated together such that the first 2,000 samples correspond to the first chain, the second 2,000 samples to the second chain and so on. In this experiment, participants completed the IGT three times with each of the three runs comprising 100 deck selections.

5.4 Experiment 3

5.4.1 Deck Selection Profiles

Experiment 3 was completed by analysing the same data presented in Experiment 1, but implemented a reduced parameter version of the EVM. For this reason, the deck selection profiles of the 25 participants are exactly the same as those presented in Section 5.2.2.

5.4.2 Posterior Distributions

The following figures are the sampled posterior distributions for the parameters of the 2-parameter EVM (in which c is held constant) for the all

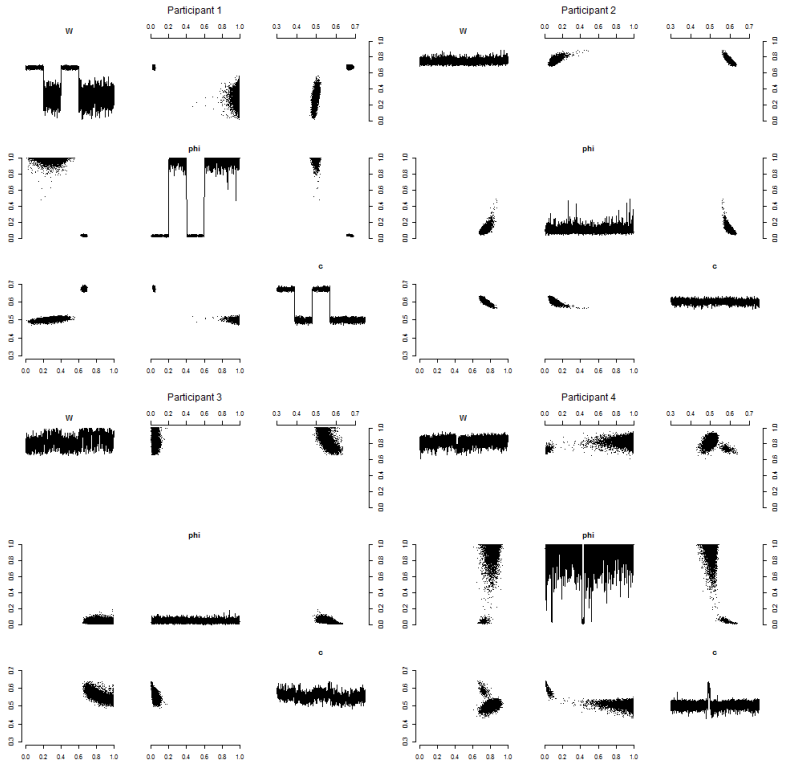

Figure 5.25: Sampled posterior distributions (with mixing) for Participants 01 through 04

.

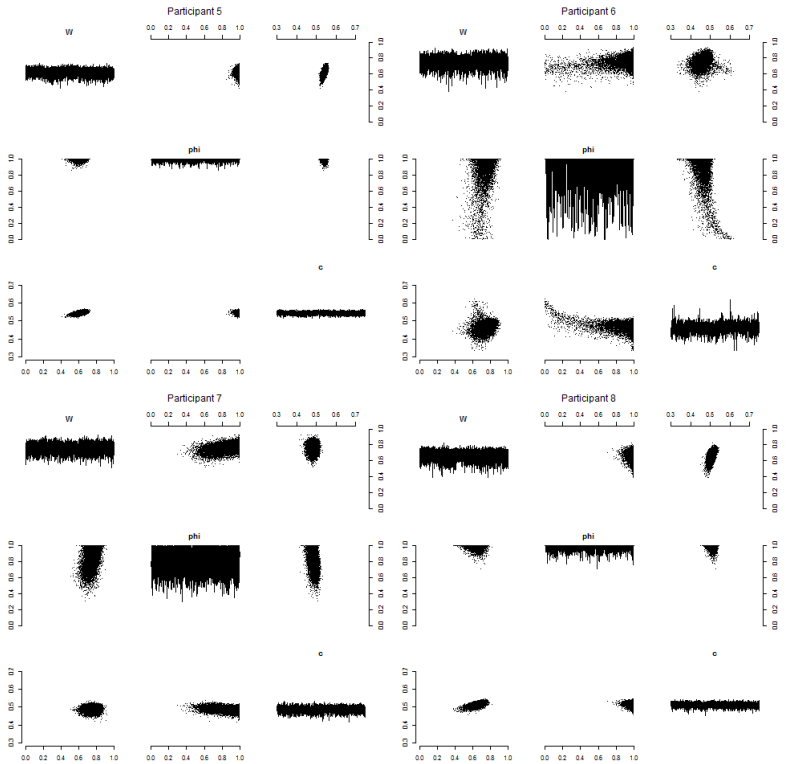

Figure 5.26: Sampled posterior distributions (with mixing) for Participants 05 through 08

.

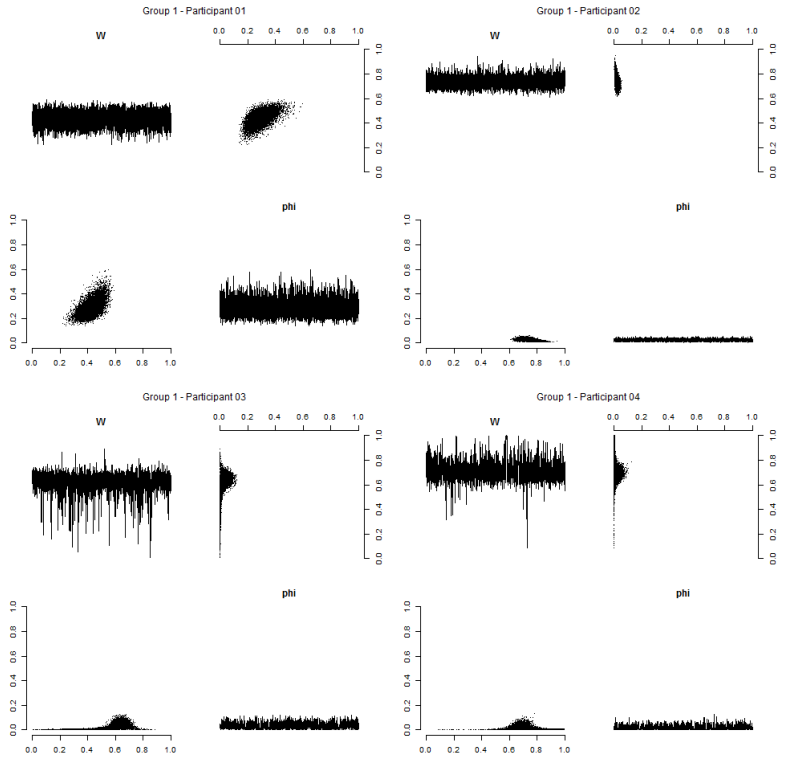

Figure 5.27: Sampled posterior distributions (with mixing) for Participants 101 to 104

.

participants completing the IGT in Experiment 1. These are pairwise plots in which the two upper diagonal panels are a reflection of the two lower diagonal panels. Across the central diagonal, the mixing for each of the two parameters (W and ϕ) are presented. This is the mixing for all five MCMC chains concatenated together such that the first 2,000 samples correspond to the first chain, the second 2,000 samples to the second chain and so on.

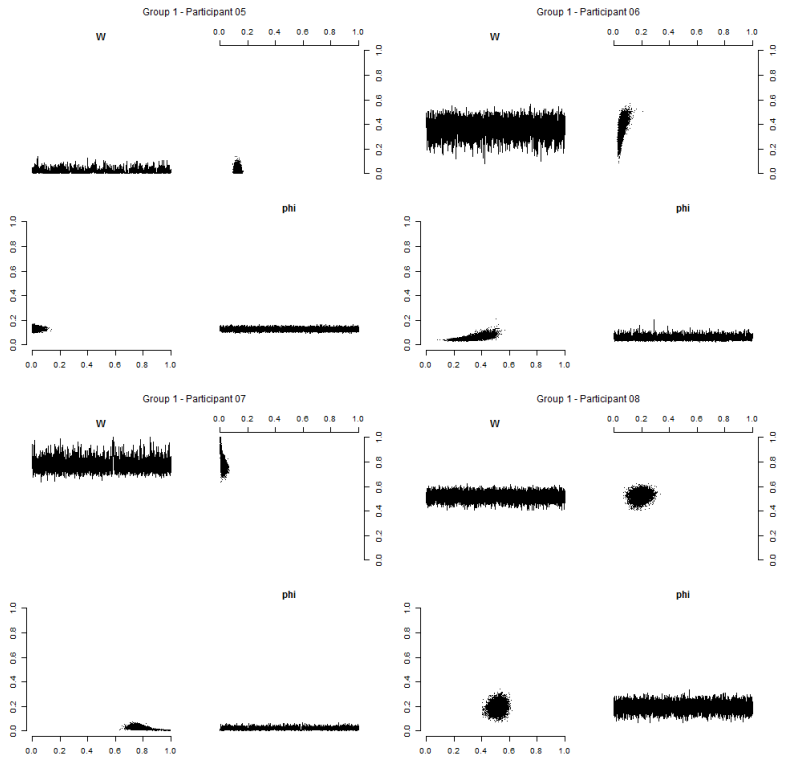

Figure 5.28: Sampled posterior distributions (with mixing) for Participants 105 to 108

.

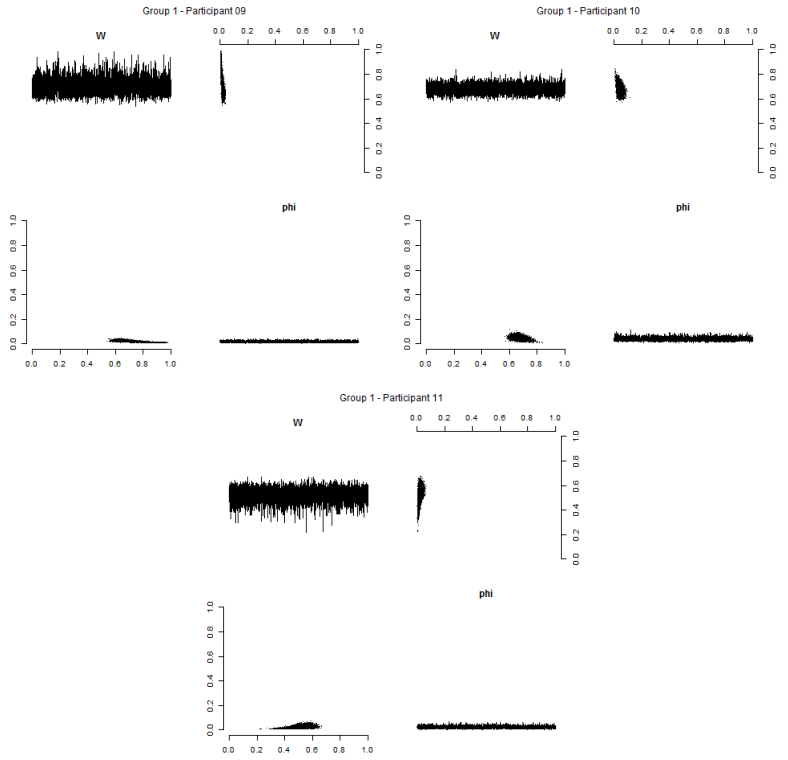

Figure 5.29: Sampled posterior distributions (with mixing) for Participants 109 to 111

.

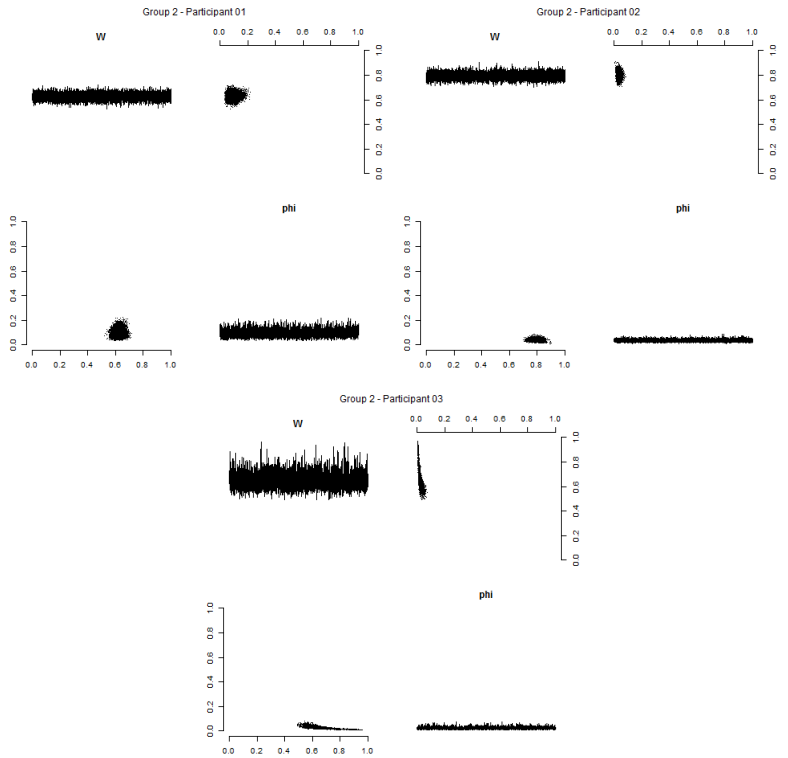

Figure 5.30: Sampled posterior distributions (with mixing) for Participants 201 to 203

.

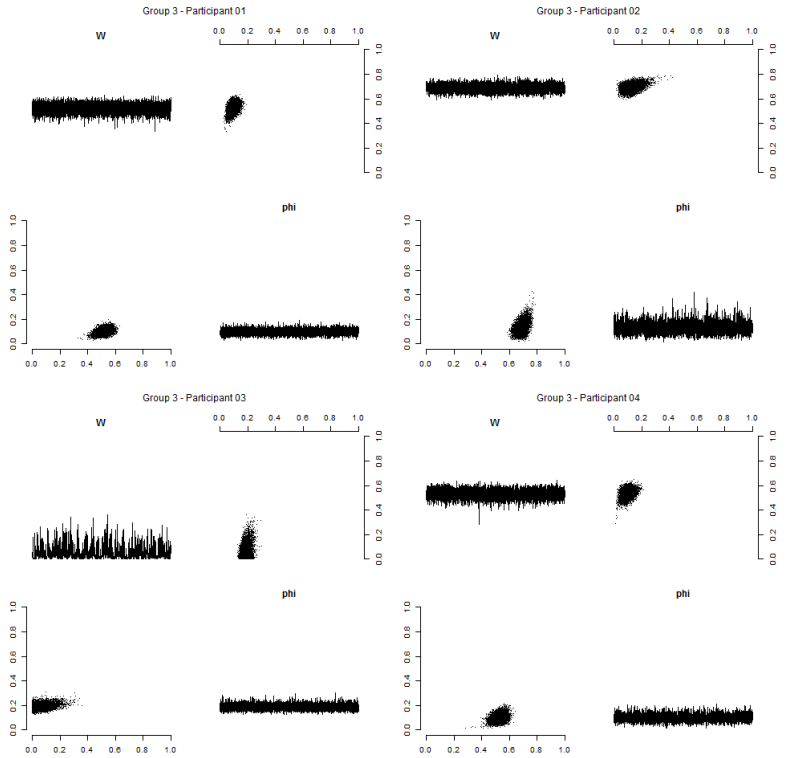

Figure 5.31: Sampled posterior distributions (with mixing) for Participants 301 to 304

.

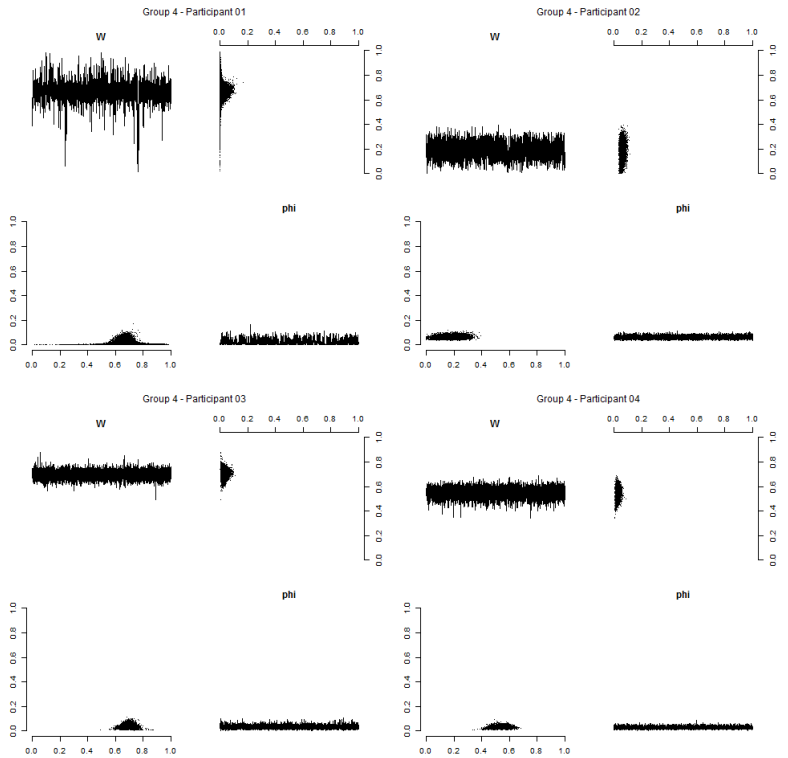

Figure 5.32: Sampled posterior distributions (with mixing) for Participants 401 to 404

.

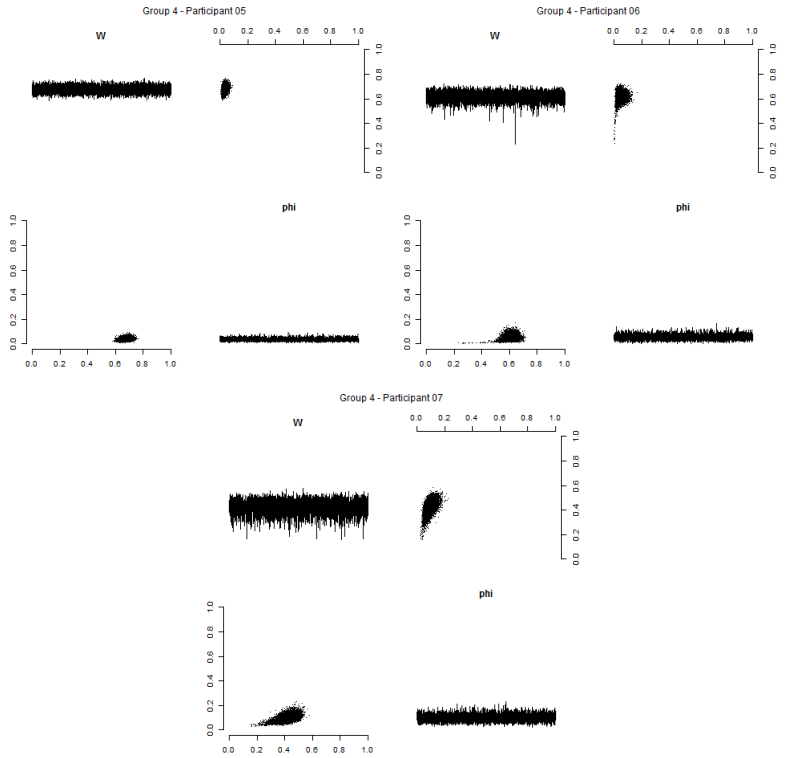

Figure 5.33: Sampled posterior distributions (with mixing) for Participants 405 to 407

.

Chapter 6

Quantifying performance on the Balloon Analogue Risk Task.

6.1 Introduction

When using any sort of model of behaviour, the goal is to link the parameters of the model with the underlying cognitive processes that give rise to observed behaviour. This is usually with the view to use the parameter estimates to either infer something about an individual's cognitive processes or gain some sort of representative, average estimate of the cognitive performance for a group of individuals. Empirical models, which are driven by the patterns in the data rather than mechanistic models of how the data has arisen, may not be able to provide any clear links back to the cognitive processes underlying observed behaviour. Mechanistic models, which are developed based on the theories surrounding the cognitive processes which give rise to observable behaviour, are therefore often favoured over their empirical counterparts. However, there is also a possibility that mechanistic models based on complex models of cognitive processes have been incorrectly specified or specified in such a way that the observed behaviour can not be described by a unique combination of the model parameters. This inability to specify parameters was exactly what we observed in Chapter 4 when considering the Expectancy Valence Model (EVM) of the Iowa Gambling

Task (IGT). The solution presented in Chapter 4 was to simplify the model, allowing for a less complex description of the individual's behaviour, but the question of whether this a true measurement of an individual's behaviour remains. Given the IGT requires such a complex system of cognitive processes to produce behaviour, it may be the case that it is just not possible to tease apart those cognitive processes by modelling the single observed behaviour of deck choice. Perhaps then, the solution is to look for a cognitive task which relies on a less complex system of cognitive processes to produce behaviour; instead of expecting one task to answer multiple questions about cognitive deficits, simplify our expectations and test one (or two) cognitive processes at a time.

The Balloon Analogue Risk Task (BART) is theorised to assess risk seeking behaviour alone (Lejuez et al., 2002), making the BART a relatively simple task compared to the IGT. In this way, modelling the cognitive processes underlying performance on the BART should be less complex than for the IGT. The following chapter will investigate whether using this cognitively less demanding task makes it easier to decompose behaviour using a less complex mechanistic model than the EVM of the IGT. The van Ravenzwaaij et al. (2011) two-parameter BART (2pB) model of behaviour will be assessed with the main aim being to ascertain if parameter estimates can be gained with enough accuracy to be useful at the level of the individual. A new, empirical model of behaviour on the BART, the Basic Response Model (BRM), is constructed based solely on the patterns in the observed data and is proposed as an alternative to the 2pB model. The BRM is then extended to a more complex model, using our understanding of participants behaviour on the task rather than our observations in the data, as a second potential alternative to the fully mechanistic 2pB model. The Run Dependent Response Model (RDRM), is therefore an example of the

fusion between mechanistic and empirical models. Whether the BRM and RDRM may have any use in modelling cognitive deficiencies at the level of the individual will also be explored.

The focus on the accuracy of the considered models at the level of the individual has particular importance in a clinical context. Clinically, the goal of psychological assessment is to ascertain if an individual may have some cognitive deficit requiring treatment or support of some kind. It is not unreasonable to think that individuals presenting for assessment may have cognitive deficits that give rise to behaviour outside of, or more extreme than, what is considered normal. In these cases the models used to measure possible cognitive deficits need to be able to accurately identify behaviours away from the normal range of results, in the extreme edges of performance. When measuring group performance, these extreme performers are labelled as outliers and their results are often ignored or averaged out in various statistical ways to lessen their impact on the overall results (see Chapter 4 for examples). However, for a model to be of clinical use, these outliers must be able to be measured as accurately as any other performer in the group. For this reason, this chapter will be a focus on the ability of the considered models to accurately identify performance away from normal, in the extreme.

6.2 The Balloon Analogue Risk Task (BART)

The BART, developed by Lejuez et al. (2002), is a computerised psychological assessment tool in which participants are presented with a virtual balloon and asked to blow it up by pressing a key on the computer. With each key press (or *pump*), the balloon increases in size and the participant wins a small amount of money. However, if the participant pumps up the balloon too far and it bursts, the participant wins nothing and moves onto

the next *trial* (another balloon). As participants have been instructed to maximise the amount of money they win, they must decide whether to keep pumping up the balloon, or whether to stop and "cash in" the balloon before it bursts and then move on to the next trial. As such, this task is designed to measure impulsivity or risk taking behaviour with impulsive individuals pumping the balloon up, more often (and observing more bursting balloons over multiple trials) than their more conservative counterparts (Hunt et al., 2005).

Unbeknown to the participants, the probability of the balloon bursting is uniformly distributed across all possible outcomes (usually the number of allowable pumps are capped to around 128, but participants do not discover this unless they pump up to the maximum) (Lejuez et al., 2002). A uniform distribution ensures that the probability of the balloon bursting is less obvious to the participant than if, for example, bursts were normally distributed about some mean. This means that, although there is technically still an expected value that participants can learn to maximise their win (ideal pump amount $= (m-1)/2$ for a maximum allowable number of pumps m), this does not become clear to an individual until many trials have been completed. As such, it is arguable that the cognitive processes associated with memory do not have the same impact on performance when completing the BART as they do on more complex tasks such as the IGT. It is the unexpected behaviour of apparently random bursting of the balloon (which is very unlike a real balloon would arguably behave in the real world) that strips back the cognitive processes and levels the playing field for those with differing memory and knowledge abilities. This is the exact thing that allows the BART to be a simpler task than the IGT and, theoretically, measure impulsivity alone. In addition, in many experiments the number of trials on the BART completed by the participants is quite low

(around 20) meaning that convergence to an "optimal" form of performance does not have time to develop. Compared to the EVM then, the BART demands a less complex system of functional cognitive processes to complete the task and, therefore, a less complex cognitive model to explain observed behaviour.

6.2.1 The Automatic BART

The original implementation of the BART by Lejuez et al. (2002), of which a simplified version is still broadly in use (see Capone et al. (2016); Lauriola et al. (2013); Peacock et al. (2013); Reynolds et al. (2006) for examples), relies on individuals pressing a button repeatedly to blow up the balloon. Several studies have found that, over relatively few trials, participants become bored with the pumping required and converge to a very low number of pumps before cashing in the balloons (Pleskac et al., 2008). When considering that the balloon bursts are uniformly distributed from $(0, m)$, where m is some number close to 130, participants who continually cash in on a very low number of pumps are unlikely to see many balloons which burst. This makes modelling extremely difficult as behaviour under loss conditions is rarely measured.

To address this problem, the Automatic BART was suggested as an alternative implementation by Pleskac et al. (2008). In the automatic BART, rather than pressing a button repeatedly to manually pump up the balloon, participants are asked to enter the number of pumps they would like to pump into a box (the *cash in* point). The balloon is then automatically pumped up until it either bursts or automatically cashes in when it reaches the number set by the participant. In this way, participants can quickly pump the balloon a large number of times reducing boredom, increasing the number of observed bursts and allowing for more accurate modelling of

behaviour (Pleskac et al., 2008). It also provides extra information over the traditional BART because you can measure the participant's *intent* on each trial. Whether the balloon bursts or not, you know how many pumps the participant intended to make. For these reasons, the current analyses are completed on the automatic BART.

6.3 Mechanistic models of behaviour

Despite the decreased cognitive processes required to complete the BART compared to tasks such as the IGT (as discussed in Section 6.2), complex cognitive models of behaviour have been proposed with Wallsten et al. (2005) suggesting a four-parameter model of behaviour. When completing the BART, an individual is required to complete several trials ($k \in (1, N)$) in which each trial consists of a balloon that needs to be pumped up until it is either cashed in or it bursts. Each trial, the participant must choose how many times they are going to pump up the balloon ($l \in (1, m)$) before cashing in. The four-parameter BART model is designed to predict the probability of the individual pumping up the balloon on trial k for each pump opportunity l. To achieve this, the four-parameter BART was designed from theoretically-based assumptions, the first of which is that each participant believes there is a probability that a pump will make the balloon burst, defined as p_k^{belief} for trial k, that is updated each trial (balloon). As the participant moves through the trials, p_k^{belief} is updated such that:

$$p_k^{belief} = 1 - \frac{\alpha + \sum_{K=0}^{k-1} n_K^{success}}{\mu + \sum_{K=0}^{k-1} n_K^{pumps}} \text{ with } \alpha > \mu. \tag{6.1}$$

Here $\sum_{K=0}^{k-1} n_K^{success}$ is the number of pumps that have not burst (successes) up to trial k, while $\sum_{K=0}^{k-1} n_K^{pumps}$ are the total number of pumps up to trial k. In this way, α provides a baseline weighting for success, while μ provides a baseline weighting for loss which means together $1 - \alpha/\mu$ can be interpreted as a measure of the participant's prior belief that the balloon will burst.

The next assumption is that the amount of pumps that the participant believes is optimal *does not change* throughout the task. Although this fact may not be relevant to the Automatic BART, the model described here utilises data from the original BART implementation in which participants are asked to manually press a button for each pump of a balloon. The number of pumps considered optimal by the participant is denoted ω_k for trial k and is expressed

$$\omega_k = \frac{-\gamma^+}{\ln\left(1 - p_k^{belief}\right)} \text{ with } \gamma^+ \geq 0. \tag{6.2}$$

This incorporates both the participant's belief that the balloon will burst (as presented in Equation 6.1), and the participant's propensity for risk taking γ^+ on trial k. Within trial k, each time a participant pumps (l) is an observation of the participant's overall decided pump amount ω_k *and* the how strongly the participant adheres to this value. The probability of the participant pumping on any pump l is therefore a probabilistic process based around ω_k. As such, the probability of pumping the l^{th} time on the k^{th} trial can be denoted

$$p_{k,l}^{pump} = \frac{1}{1 + \exp\beta\left(l - \omega_k\right)} \text{ with } \beta \geq 0 \tag{6.3}$$

where β is a measure of the participant's behavioural consistency. These assumptions were modelled in such a way as to provide information about an individual's baseline weighting for the likelihood of success and loss, the participant's propensity for risk taking and the participant's behavioural consistency.

van Ravenzwaaij et al. (2011), however, showed that the four-parameter model proposed by Wallsten et al. (2005) was unable to be solved uniquely in some contexts. Much like the problems described in Chapter 4 for the EVM of the IGT, van Ravenzwaaij et al. (2011) showed the four-parameter BART models complexity resulted in non-unique estimates of the α and μ

parameters that had potentially conflicting cognitive interpretations. van Ravenzwaaij et al. (2011) proposed that a two-parameter version of the model was more successful.

6.3.1 The Two-parameter van Ravenzwaaij Model

In the two-parameter van Ravenzwaaij et al. (2011) model Equation 6.1, probability p_k^{belief}, is fixed across all trials. Given that $p_k^{beleif} \in (0, 1)$, Equation 6.2 can now be written as:

$$\omega_k = \alpha, \text{ with } \alpha \geq 0 \tag{6.4}$$

where α is a rescaled version of γ^+. This means Equation 6.3 is now

$$p_{k,l}^{pump} = \frac{1}{1 + \exp \beta \left(l - \alpha\right)}. \tag{6.5}$$

So the two-parameter model of the BART (2pB) relies on two parameters α and β where $\beta \in (0, 1)$ and, with balloon pumps capped at m pumps, $\alpha \in (0, m)$. The parameter α then represents the point at, above which, the individual believes the balloon is more likely to burst than not. For example, a participant with $\alpha = 40$ would be more likely to cash in their balloon post-40 trials ($l > 40$) than before. To measure the individual's conviction in the α cut-off value, β is included to either soften or strengthen the rate at which the probability of pumping changes across pumps (the participant's behavioural consistency). For example, taking our individual with $\alpha = 40$, if they also had $\beta = 0.05$ that would suggest they were not confident in this cut-off value and their probability of cashing in would not change greatly from $l = 38$ to $l = 42$. A $\beta = 0.9$ alternately would see an immediate shift from a probability of cashing close to zero, to a probability of cashing close to one over the same interval.

6.4 Empirical models of behaviour

As an alternative to the two-parameter mechanistic model (2pB) proposed in Section 6.3, two empirical models of performance on the BART were also developed and are presented in this Chapter.

Empirical models are formed, instead of being based on the cognitive theory underlying behaviour, by being led by the data itself. A model is first created to explain the observed data and then links are drawn between the fitted parameters and the potential cognitive processes they may represent. Importantly, the links to cognitive processes are drawn *after* the model has been decided. As discussed in Section 3.1, although this may increase the ability of finding a model that is identifiable, this does not guarantee a model that will link well to the cognitive processes which give rise to the observed behaviour.

6.4.1 Basic Response Model

Starting with the most basic, one step ahead prediction model, the Basic Response Model (BRM) is based on data from Automatic BART tasks. Led entirely by the data, rather than the cognitive processes underlying it, it seeks to explain increases and decreases in the number of pumps guessed based on the events that the participant has experienced.

Assuming that the actual number of pumps guessed on trial $k+1$, $Pump_{k+1}$ will be normally distributed around some expected value μ with some standard deviation σ, we can expect

$$Pump_{k+1} \sim \mathrm{N}\left(\mu_{k+1}, \sigma\right). \tag{6.6}$$

If we consider the prediction for the expected value μ_{k+1} of the number of pumps on trial $k+1$, we get

$$\mu_{k+1} = Pump_k + \delta_k D + (1 - \delta_k) I \tag{6.7}$$

where $Pump_k$ is the number of pumps actually guessed on trial k and

$$\delta_k = \begin{cases} 1, \text{ if } Pump_k \text{ resulted in a burst} \\ 0, \text{ if } Pump_k \text{ did NOT result in a burst.} \end{cases}$$

In this way

$D < 0$ shows a decrease in number of pumps after a burst

$D = 0$ shows no change in number of pumps after a burst

$D > 0$ shows an increase in number of pumps after a burst

while

$I < 0$ shows decreases in pumps after a success

$I = 0$ shows consistent number of pumps after a success

$I > 0$ shows an increase in number of pumps after a success.

The parameter D empirically models the change to the number of pumps chosen following unsuccessful results. However, a clear link between D and psychological behaviour is also clear as it measures how a loss, or bursting balloon, effects behaviour. More risk averse individuals should decrease their number of pumps following a burst while impulsive individuals may increase their pump guess. As such, D can be seen as measuring risk aversion or impulsivity.

Also measuring increases and decreases in response to the outcomes of the task, I is concerned with behaviour following wins (or successful cash ins). Similarly to the D parameter, I also has clear links to psychological behaviour with participants who gain great confidence from a win more likely to increase subsequent pump guesses by a large amount whereas a lack of confidence may see decreases even in face of a win.

With D representing impulsivity following a loss and I confidence after a gain, the parameter σ measures how closely an individual's pump guesses

are to those estimates. Small values of σ models increases and decreases to pump guesses that are close to the preferred values of D and I, where as large values of σ allows more variability in pump guesses on each trial. In this way, σ can be thought of as a measure of behavioural consistency.

Of course, the BRM does assume that both D and I are constant across all trials and yet it is realistic to think this may not be a fair assumption. As a participant on the task encounters sequential trials where the balloon doesn't burst, it would be reasonable to think their behaviour may change. As such, a more complex model that models this behaviour was considered.

6.5 Where Empirical and Mechanistic Models Meet

The Basic Response Model (BRM) is essentially summarising the most simple empirical features of the BART data that are presented; how big the subsequent increases and decreases are in response to observed trial outcomes. It is not hard to imagine, however, that the choices a participant makes when completing the task are more complex than a simple increase or decrease rule. You can imagine that, as the number of trials without a burst increases, an individual may become more confident and take bigger risks. For others, an increased run of wins may result in more cautious increases in expectation of a burst. The next model we suggest then, the Run Dependent Response Model (RDRM), takes the BRM and incorporates a feature designed to capture some of this increased complexity. As such, we have an empirical model which has been extended to include a meausre of our theories surrounding observed behaviour, which are mechanistic in nature. So, in the RDRM we see one example of where empirical and mechanistic model may converge.

6.5.1 Run Dependent Response Model

The Run Dependent Response Model (RDRM), like the BRM, uses one step ahead predictions to estimate the pump amount on trial $k+1$, based on the results of the previous trials. However, it also takes into account if there has been a *run* R_k of wins or successful cash ins in a row. The number of pumps predicted for trial $k+1$ is therefore described as

$$\mu_{k+1} = Pump_k + \delta_k D + (1 - \delta_k)\, I R_k^{\phi} \tag{6.8}$$

where, just as in the BRM,

$$\delta_k = \begin{cases} 1, \text{ if } Pump_k \text{ resulted in a burst} \\ 0, \text{ if } Pump_k \text{ did NOT result in a burst} \end{cases}$$

and D and I are as described for the BRM in Section 6.4.1. The difference between the BRM and the RDRM is that the parameter I is now multiplied by R_k, which is the run or the number of trials since the last burst was observed. The run, R_k has power ϕ which can take the following values:

$\phi < 0$ shows decreases in magnitude of pump guesses as run length increases

$\phi = 0$ shows a constant magnitude of pump guesses, independent of run length (this is the BRM)

$\phi > 0$ shows an increase in magnitude of pump guesses as run length increases

$\phi = 1$ shows a linear increase in magnitude of pump guesses as run length increases.

For the RDRM then, although I is still concerned with modelling behaviour following wins (or successful cash ins), the addition of the parameter ϕ attached to information about the run of successful trials enables this behaviour to be more trial dependent.

By taking into account the number of successful trials in a row, and modelling the impact of that on future guesses, the level of confidence experienced by the participant can now change across trials. In this way, the ϕ parameter should enable us to tease apart individuals whose confidence increases following sequential wins from those who become more cautious by multiplying the constant I by the transformed R_k.

6.6 Method

Parameter recovery simulation studies were conducted separately for each of the considered models of behaviour on the BART. Data was simulated across reasonable values (defined separately below) for each of the considered models and then, using the same model which generated the data, it was ascertained if the model parameters could be recovered with any level of accuracy.

All statistical analysis was conducted using the statistical software R (The R Development Core Team, 2009), and the code for all computations is available by contacting the author directly.

Two-parameter BART (2pB)

In their simulation studies, van Ravenzwaaij et al. (2011) used $\beta = 0.4, 0.5, 0.6, 0.7, 0.8$ and $\gamma^+ = 0.6, 1, 1.4, 1.8, 2.2$ as the values for which data was generated. Given that we are particularly interested in looking at the extremes of behaviour, $\beta = 0.05, 0.2, 0.4, 0.6, 0.8, 0.95$ was considered and, given that α is a rescaled version of γ^+ (Section 6.3.1, Equation 6.4) where

$$\alpha = -\gamma^+ \frac{1}{\ln(a)}, a \in (0,1),$$

we considered $\alpha = 5, 10, 20, 50, 80, 95$.

For each possible combination of α and β, data was simulated for 100 synthetic participants using the 2pB model proposed by van Ravenzwaaij

et al. (2011) (Equation 6.5). Each synthetic participant completed a 250 trial version of the BART with a maximum allowable pump guess of 130 pumps. Once the data was simulated, optimisation of the negative log likelihood surface (using a Nelder-Mead optimisation algorithm in the optim function in R) was completed and the parameter estimates for α and β stored. The log likelihood was generated for across trials k and pump opportunities l for all trials n_k and all pump opportunities within trials $n_{l(k)}$ based on the probability of data D:

$$\mathrm{P}\left(D|\alpha,\beta\right) = \prod_{k=1}^{n_k} \prod_{l=1}^{n_{l(k)}} p_{k,l}^{pump} \left(1 - p_{k,n_{l(k)}+1}^{pump}\right)^{d_k} \tag{6.9}$$

where $d_k = 0$ if the balloon burst on trial k, $d_k = 0$ otherwise and $p_{k,l}^{pump}$ is as defined in Equation 6.5. The distributions of the parameter estimates for each parameter combination were then analysed and are presented in Section 6.7.1.

Basic Response Model (BRM)

Simulation values considered for the BRM included every combination of $D = -20, -15, -10, -5, 0$ with $I = 0, 5, 10, 15, 20$. For each unique combination (25 in all), 100 synthetic participants were created simulating responses over a 100 trial version of the Automatic BART (maximum allowable pump of 130). Each pump guess then was simulated using a normal distribution with mean defined by the BRM (Equation 6.7) and standard deviation generated at three levels: $\sigma = 1, 5, 10$.

Once the data was simulated, optimisation of the negative log likelihood surface (using a Nelder-Mead optimisation algorithm in the optim function in R) was completed and the parameter estimates for D, I and σ recovered.

Rearranging Equation 6.7, we can see that to estimate parameters D

and I we need

$$\mu_{k+1} - Pump_k = \delta_k D + (1 - \delta_k) I.$$

Of course, we don't observe μ_{k+1} directly, but this can be approximated with

$$Pump_{k+1} - Pump_k \approx \delta_k D + (1 - \delta_k) I. \tag{6.10}$$

Here we are estimating the *change* ($Change_{k+1}$) between pumps $k+1$ and k with either D on trials where the balloon burst or with I when the balloon was successfully cashed in. In this way, the likelihood can be completely decoupled with estimates for D and I reliant on non-overlapping results.

Although the observed pumps are discrete in nature, there are 99 difference scores for each generated data set used to estimate the parameters in the BRM (100 trials per simulated participant). Given the large N for each optimisation, and that a normal distribution was used to simulate the data, it is fair use a normal distribution to model the observed distribution of difference scores.

So to estimate the parameters D, I and σ, we maximise the negative log likelihood

$$\begin{aligned} -\log(\mathrm{L}(D, I, \sigma | Change)) \quad &= \\ &- \sum_{k=1}^{N-1} \delta_k \log(\mathrm{P}(Change_{k+1} | D, \sigma^2)) \\ &- \sum_{k=1}^{N-1} (1 - \delta_k) \log(\mathrm{P}(Change_{k+1} | I, \sigma^2)) \end{aligned} \tag{6.11}$$

where P is the cumulative probability of the normal distribution and N is the total number of trials.

The distribution of the parameter estimates for each parameter combination of D, I and σ were then analysed and are presented in Section 6.7.1.

Run Dependent Response Model (RDRM)

Simulation values considered for the RDRM included every combination of $D = -15, -10, -5, 0$ with $I = 0, 5, 10, 15$ and $\phi = -1, 0, 0.5, 1.5$. For each unique combination (64 in all), 100 synthetic participants were created simulating responses over a 100 trial version of the Automatic BART (maximum allowable pump of 130). Each pump guess then was simulated using a normal distribution with mean defined by the RDRM (Equation 6.8) and standard deviation generated at three levels: $\sigma = 1, 5, 10$.

Once the data was simulated, optimisation of the negative log likelihood surface (using a Nelder-Mead optimisation algorithm in the optim function in R) was completed and the parameter estimates for D, I and ϕ recovered. The distribution of the parameter estimates for each parameter combination were then analysed and are present in Section 6.7.1.

Rearranging Equation 6.8, we can see that to estimate parameters D, I and ϕ we need

$$\mu_{k+1} - Pump_k = \delta_k D + (1 - \delta_k)\, IR_k^{\phi}.$$

Similarly to Equation 6.10, we don't observe μ_{k+1} directly, but this can be approximated with

$$Pump_{k+1} - Pump_k \approx \delta_k D + (1 - \delta_k)\, IR_k^{\phi}. \tag{6.12}$$

However, unlike the BRM, the likelihood for the RDRM can not be completely decoupled due to the relationship between I and ϕ. So, assuming the differences can be approximated using a normal distribution (as described for the BRM), the Change observed for trial $k+1$ can be estimated as using

$$Change_{k+1} \sim N\left(\delta_k D + (1 - \delta_k)\, IR_k^{\phi}, \sigma\right)$$

So to estimate the parameters D, I, ϕ and σ, we maximise the negative

log likelihood

$$
\begin{aligned}
-\log(\mathrm{L}(D, I, \phi|Change)) &= \\
&- \sum_{k=1}^{N-1} \delta_k \log(\mathrm{P}(Change_{k+1}|D, I, \phi, \sigma^2))
\end{aligned}
\tag{6.13}
$$

where P is the cumulative probability of the normal distribution and N is the total number of trials.

The distribution of the parameter estimates for each parameter combination were then analysed and are presented in Section 6.7.1.

6.6.1 Model Comparison

Because the 2pB is not nested within the BRM or RDRM, comparison between the model fits is not straight forward. In an article by Wagenmakers et al. (2004), a process is described for comparing non-nested models such as these making use of the Goodness of Fit statistics (here the AIC is used) from parametric bootstrap samples. This is termed *model mimicry* and looks at how well different models can explain the variability in data generated using the model itself and the models it is being compared to. In the process suggested by Wagenmakers et al. (2004), they began with a non-parametric bootstrap from the observed data. However, given that the data in this study was entirely simulated, the following process begins with data simulated across a full range of parameter values:

Model Mimicry Stages

1. Fit both model A and B to the simulated data and gain parameter estimates θ_A and θ_B, Goodness of Fit (GOF) values GOF_A and GOF_B and the difference in GOF ΔGOF_{AB}

    - ΔGOF_{AB} is $GOF_A - GOF_B$. The difference in AIC between the models A and B when fit to the original data.

2. Generate data X_A with model A using parameter estimates θ_A and generate data X_B using model B using parameter estimates θ_B

3. Fit both model A and model B to both X_A and X_B gaining estimates $\theta_{X_A,A}$, $\theta_{X_A,B}$ and $\theta_{X_B,A}$, $\theta_{X_B,B}$

4. Ascertain Goodness of Fit (GOF) statistics for each of the fits $GOF_{X_A,A}$, $GOF_{X_A,B}$ and $GOF_{X_B,A}$, $GOF_{X_B,B}$

5. Look at the difference between GOF for the samples $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$

    - $\Delta GOF_{X_A,AB}$ is $GOF_{X_A,A} - GOF_{X_A,B}$. The difference in AIC between models A and B when fit to the data simulated using $\theta_{X_A,A}$ and $\theta_{X_A,B}$.

This is then repeated many times for each of the possible combinations across the three models, resulting in a distribution of differences in the GOF values. If the distribution of values is mainly negative, then model A must be preferred, where model B is preferred if the distribution is positive. If the models can mimic each other well, there may be no clear preference between models (the distribution is centred around zero) but it would be expected that there should be some preference toward the model which simulated the data. For example, $\Delta GOF_{X_A,AB}$ should be negative, suggesting a preference for model A as this is the model that created the data here.

The original GOF values can also be compared to the simulated distributions to ascertain which model is fitting the data best in which scenario. Figure 6.1 shows an example of how the original data may be classified using the differences in GOF distributions. Looking at the distributions of $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ concurrently will allow the overlap between these distributions to be measured. Ascertaining where the original data

ΔGOF_{AB} falls in relation to this distribution allows for cross classification of the original data. If the data was created using process A, then the ΔGOF_{AB} should fall inside the distribution for $\Delta GOF_{X_A,AB}$ and, should there be enough distinction between $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$, this will allow for it to be correctly identified as being created using model A. If there is too much overlap between $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ however, the ability to classify ΔGOF_{AB} will be lost.

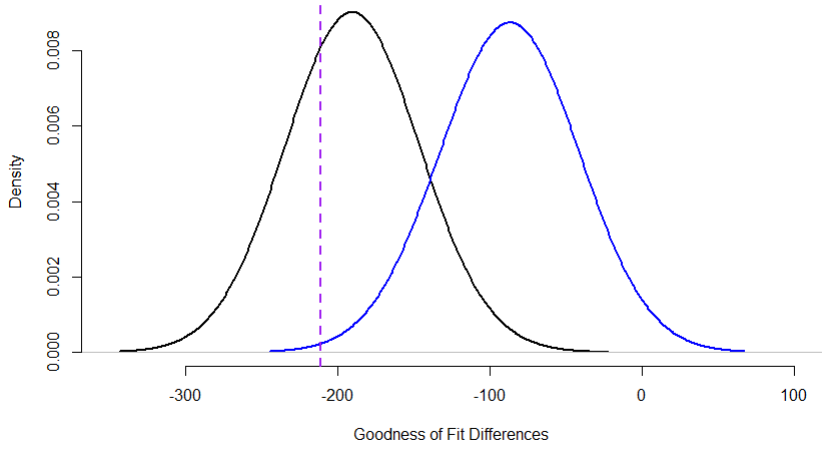

Figure 6.1: The probability densities for a hypothetical $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$. On the left is the ΔGOF_{AB} from the data generated using model A and the ΔGOF_{AB} from the data generated using model B is on the right. If model A is true, then the ΔGOF_{AB} from the original data should fall within the left distribution while, if model B is true, it should fall in the right. The purple dotted line represents an example of a ΔGOF_{AB} that would suggest model A is preferred.

Wagenmakers et al. (2004) also showed that, using the standard deviation of the distributions allows the formation of a selection rule ascertaining the distribution from which the ΔGOF_{AB} may have come from. For $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ we obtain means μ_A and μ_B and standard deviations σ_A and σ_B respectively. Assuming here that $\mu_A < \mu_B$ and $\mu_A - 2*\sigma_A < \mu_B - 2*\sigma_B$ then if x^* (an individual observation from ΔGOF_{AB}) is less than $\mu_B - 2*\sigma_B$ then the data is preferring Model A. If $x^* > \mu_A + 2*\sigma_A$ then Model B is preferred. In the case that $\mu_A - 2*\sigma_A \geq \mu_B - 2*\sigma_B$, then ei-

ther there is too much similarity between the $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ distributions and x^* can not be classified, or x^* can be classified as preferring Model A and never B due to the nesting of the distribution of B within A.

In the current implementation, the 25 parameter combinations across the values $D = -20, -15, -10, -5, 0$ with $I = 0, 5, 10, 15, 20$ for the BRM were used to simulate one set of base data while 36 parameter combinations across $\alpha = 5, 10, 20, 50, 80, 95$ and $\beta = 0.05, 0.2, 0.4, 0.6, 0.8, 0.95$ for the 2pB were to simulate the second base set of data. Twenty data sets were simulated for each parameter combination for each set of data (500 and 720 in total respectively). The two models, 2pB and BRM were then fitted to each set of the base data, parameter values stored and a further 20 data sets simulated for each of the parameter values obtained (10,000 and 14,400 data sets in all, respectively). The parameter values were then obtained for the 10,000 and 14,400 data sets and the AIC calculated for each fit using the formula $AIC = 2k - 2 \times \log$ likelihood. The values of k used were as follows:

$$k = \begin{cases} 2, & \text{2pB } (\alpha, \beta) \\ 3, & \text{BRM } (D, I, \sigma) \end{cases} \tag{6.14}$$

The difference between AIC values were then calculated and compared across the 25 groups of 400 similar samples for the BRM base data and across the 36 groups of 400 similar samples for the 2pB base data.

6.7 Results

6.7.1 Parameter Recovery

Two-parameter BART (2pB)

Before analysing any data for the 2pB van Ravenzwaaij et al. (2011) model, the relationship between the parameters α and β were investigated. As a

reminder, from Equation 6.5, we are estimating α and β from the equation

$$p_{k,l}^{pump} = \frac{1}{1 + \exp \beta \, (l - \alpha)}$$

where l is the current pump ($l \in (1, 130)$) on trial k. Figure 6.2 shows the relationship between α and the probability of cashing in for different values of β. With the current pump set to $l = 50$, it is clear from Figure 6.2 that moving relatively small distances from l results in dramatic changes in probability for many values of β. In fact, any $\beta > 0.5$ is hard to distinguish from one another at α values only ± 10 away from l.

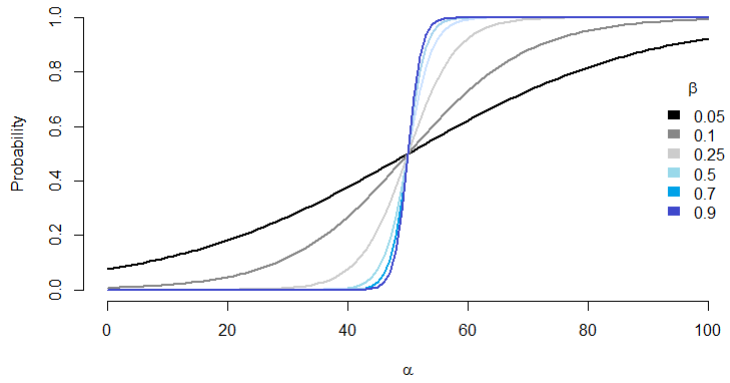

Figure 6.2: The relationship between α and the probability of cashing in for different estimates of β on the 2pB van Ravenzwaaij et al. (2011) model. The current pump $l = 50$ is set for illustrative purposes.

Consider a situation where an individual has $\alpha = 70$. This is not a particularly extreme value when the number of pumps l on any given trial k have a maximum of $m = 130$. Yet Figure 6.3 shows that, in this case, it would be extremely difficult to identify the difference between β values greater than 0.25. Irrespective of whether l is set to 50 or 70, when β is small, the probability profiles become so close to one another that the difference between them is hard to see. Once an individuals random error is added in, this will make it incredibly difficult to ascertain the correct estimate for β, irrespective of well α is estimated. Similarly to the demonstrated $\alpha = 70$, $\beta > 0.25$ is not considered an extreme value and so this situation is

one in which an individual with parameters within the range of reasonable consideration may be hard to identify.

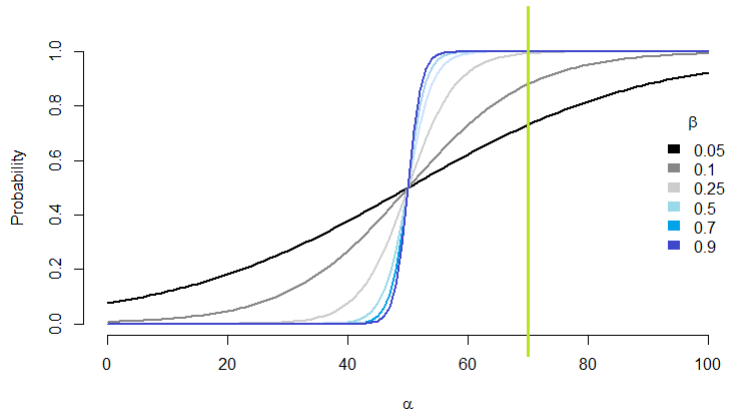

Figure 6.3: The relationship between α and the probability of cashing in for different estimates of β on the 2pB van Ravenzwaaij et al. (2011) model. The current pump $l = 50$ is set for illustrative purposes. A line is drawn at the $\alpha = 70$ line to highlight the increasing difficulty in differentiating between the different β values as values of α move further from l. This makes estimation of β more difficult.

In addition to the potential inability to identify β for a large range of α values, Figure 6.4 displays the probability of cashing in for an estimate of $\beta = 0$. The flatness observed in the $\beta = 0$ line shows an equal probability of cashing in for any pump l for trial k making it impossible to tell which α value the individual may have. Although $\beta = 0$ is truly an extreme value, Figure 6.5 highlights that there also may be identifiability problems of β.

When β takes values < 0.3 or > 0.8, Figure 6.5 shows it becomes increasingly difficult to distinguish between the lines for $\alpha = 10$, 30, 45 or $\alpha = 55$, 70, 90. Values of $\beta > 0.8$ may be considered somewhat extreme, but values of $0.3 < \beta < 0.8$ are well within a reasonable range of expected values. Couple that with the reasonable values of $\alpha = 30$, 45, 55, 70 and these plots again highlight a potential problem. Reasonable values of the parameters in the 2pB van Ravenzwaaij et al. (2011) model may be hard to identify precisely.

The parameter recovery simulations confirmed that there may be a prob-

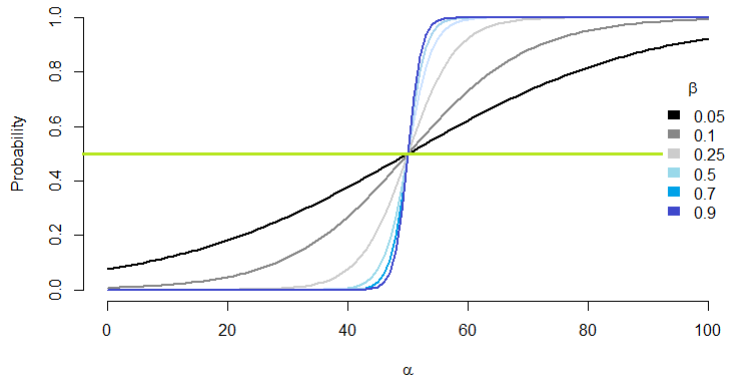

Figure 6.4: The relationship between α and the probability of cashing in for different estimates of β on the 2pB van Ravenzwaaij et al. (2011) model. The current pump $l = 50$ is set for illustrative purposes. A line is drawn at the $\beta = 0$ line to highlight that very low values of β can flatten the probability to a point where it is hard to distinguish where l resides. This in turn makes it harder to estimate α.

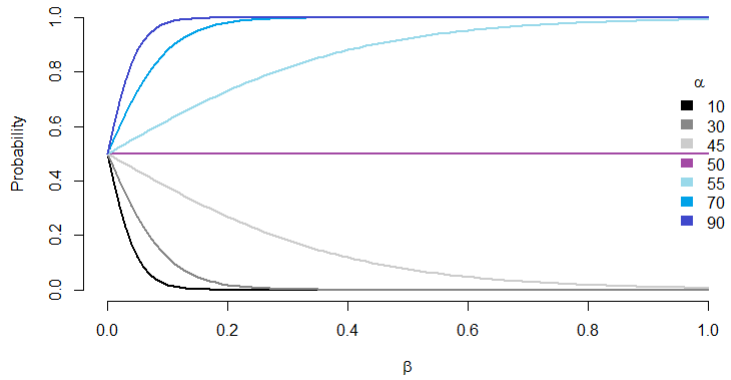

Figure 6.5: The relationship between β and the probability of cashing in for different estimates of α on the 2pB van Ravenzwaaij et al. (2011) model. The current pump $l = 50$ is set for illustrative purposes. For high values of β, it is clear that large or small values of α become harder to distinguish.

lem with identifiability in this model with both α and β hard to recover for certain values. Figure 6.6 shows that it is an interaction between the values of the two parameters that make the recovery of parameter values less accurate for combinations of α and β. Specifically, for simulated values of $\alpha \geq 50$, as parameter values of β increased the recovery of β became less accurate with some estimates more than 0.35 different to the value used to simulate the data.

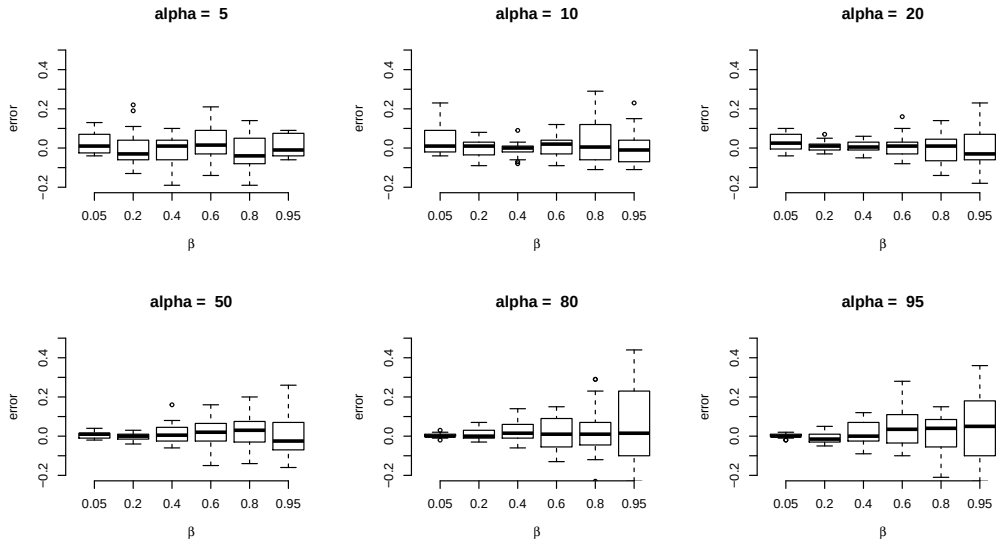

Figure 6.6: The pattern of change in the estimates of β across the imputed α values for the two parameter van Ravenzwaaij model.

Table 6.1: Number of converged maximum likelihood estimates recovered for pairs of α and β estimates on the 2pB van Ravenzwaaij et al. (2011) model.

	α					
β	5	10	20	50	80	95
0.05	15	17	20	23	25	25
0.2	17	23	21	24	24	24
0.4	22	18	22	24	24	24
0.6	22	12	24	24	24	24
0.8	14	14	24	24	24	24
0.95	8	17	24	24	24	24

Figure 6.7 highlights that α was generally recovered more accurately than β, except when β values were small. In this case, the α estimates became increasingly more variable, with some estimates 20 points different to the value used to simulate the data, highlighting the uncertainty surrounding the estimate. Given the relationship between α and $\beta \approx 0$ displayed in Figure 6.5, this lack of identifiability around small β was expected.

Basic Response Model (BRM)

Before analysing the results of the BRM, the viability of the generated data first had to be ascertained. When asking an individual to complete a psychological assessment tool, if the individual completes the task in such a

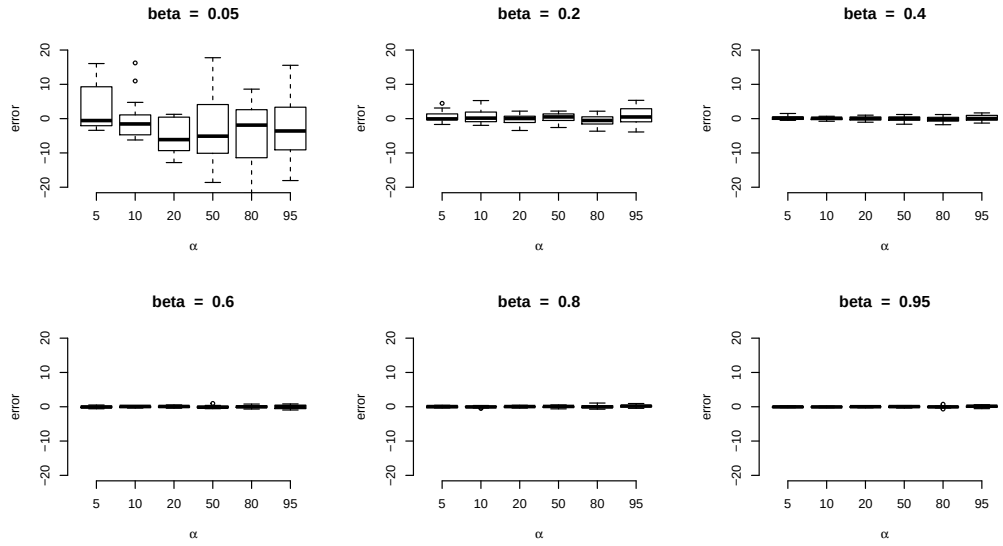

Figure 6.7: The pattern of change in the estimates of α across the imputed β values for the two parameter van Ravenzwaaij model. The y-axis here has been constrained to be between -20 and 20 to better illustrate variability at smaller intervals. That means some outliers are not visible in the $\beta = 0.05$ plot.

way that viable data can not be obtained, it is often excluded from analyses. In the case of the generated data if an individual run did not see at least five balloons burst or cash in successfully at least five times, then the data was excluded from analyses. Practically this would reflect an individual who cashes in their pumps on the first pump for a very large number of trials or who guesses the maximum number on almost every trial. In both of these cases, it would be assumed that either the participant did not understand the instructions of the task, did not want to participate or has some extreme cognitive reason for their behaviour which warrants further investigation outside of the task.

Table 6.2 shows the number of simulated data sets which were considered viable (contained at least five each of balloon bursts and successful cash-ins) for each parameter combination of D and I. The only times that less than 100% of the simulated data sets were viable occurred when either $D = 0$ or $I = 0$. Intuitively, given that $D = 0$ results in no change to the pump amount guessed following a decrease and, alternately, $I = 0$ results in no

Table 6.2: Number of viable data sets available for analysis for pairs of D and I estimates on the BRM (with standard deviation $\sigma = 1$, 5 and 10). Non-viable data sets were only observed in the cases where either $D = 0$ or $I = 0$ and, in all but five cases, greater than 95% of simulated data sets were viable.

	$\sigma = 1$					$\sigma = 5$					$\sigma = 10$				
						I									
D	0	5	10	15	20	0	5	10	15	20	0	5	10	15	20
-20	88	100	100	100	100	79	100	100	100	100	97	100	100	100	100
-15	89	100	100	100	100	87	100	100	100	100	98	100	100	100	100
-10	98	100	100	100	100	88	100	100	100	100	98	100	100	100	100
-5	100	100	100	100	100	98	100	100	100	100	100	100	100	100	100
0	100	100	99	100	97	95	99	97	99	95	100	100	100	99	100

change following a successful cash in, this result is not unexpected. In both of these cases we would expect to see convergence to the maximal or minimal allowable pump amount, more so if the corresponding I or D value is quite large. Couple this with the fact that the initial pump guess $Pump_1$ was drawn from a uniform distribution $Pump_1 \sim U(1, 130)$ it is quite possible that some simulated data sets would start close to the maximum or minimum allowable pump amount and then converge quite quickly to the maximum or minimum. This, of course, would result in very few bursts or very few successful cash-ins dependant on the combinations of parameters and an associated decrease in the number of viable data sets.

Given that only five parameter combinations observed in Table 6.2 had less than 95% of data sets returned as viable, and that the lowest number of viable data sets was 79 ($D = -20$, $I = 0$, $\sigma = 5$), there was a large enough sample size for each parameter combination for analyses to be completed.

When considering the difference between the imputed D and I values and those that were recovered across the simulated data, Figure 6.8 (when $\sigma = 1$) shows the distribution of the differences tightly clustered around zero suggesting accurate parameter recover in the majority of cases. However, despite the generally tight clustering around zero, both histograms in Figure 6.8 are skewed with some recovered parameters up to 10 units different to the value used to create the data.

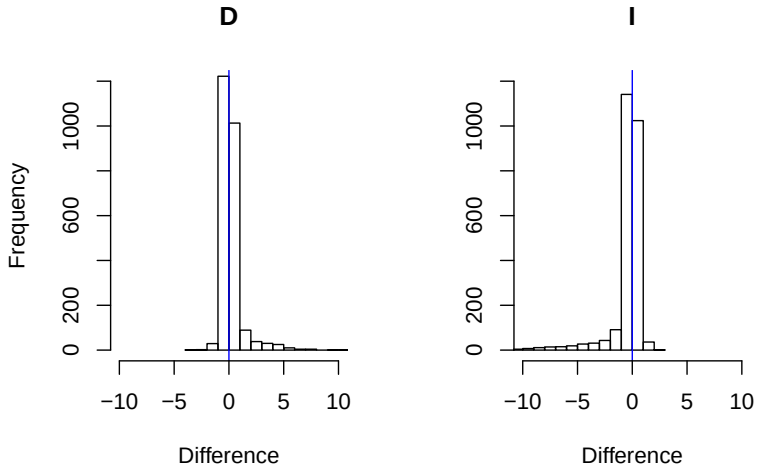

Figure 6.8: The difference between the recovered D and I values and the inferred values for the BRM (with standard deviation $\sigma = 1$). The majority of all observations (98.1%) were recovered within ± 5 units with a large peak in Difference values either side of zero.

Despite the skewed results, Figure 6.8 shows that 99.2% (2449/2469) of the D values and 97.0% (2396/2469) of the I values were recovered within ± 5 units. These high numbers suggest that, despite a few skewed results, there was a reasonable level of accuracy in parameter recovery in general.

The skewed results of the differences observed in Figure 6.8 are also observable in Figure 6.9, where $\sigma = 5$, but with the increased σ value the amount of skew appears less extreme. The tight peak observed around zero in Figure 6.8 has softened in Figure 6.9 and the histogram is appearing more normally distributed. This pattern continues in Figure 6.10 where the histograms look much more normally distributed and the skew is far less extreme.

Similarly to the results for $\sigma = 1$, when $\sigma = 5$ or $\sigma = 10$ there was still a reasonable level of parameter recovery in general. When $\sigma = 5$, D was recovered within ± 5 units 98.8% (2408/2437) of the time with I recovered 98.0% (2389/2437) of the time. Similarly, when $\sigma = 10$, D was recovered 99.0% (2467/2492) and I was recovered 98.3% (2449/2492) of the time within ± 5 units. This suggests that the majority of the time,

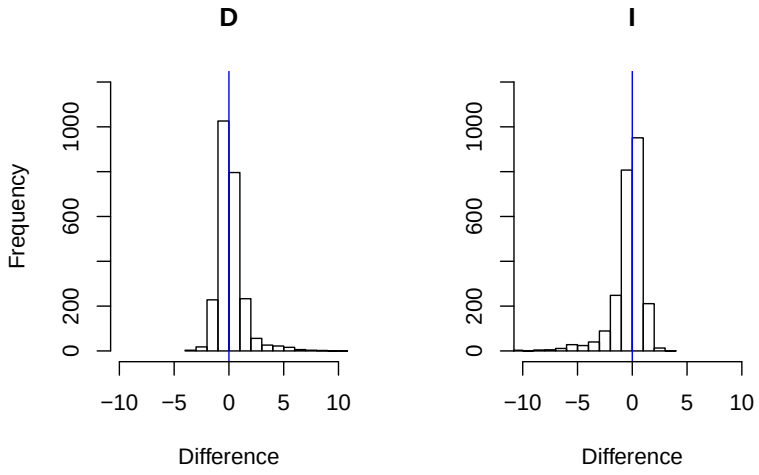

Figure 6.9: The difference between the recovered D and I values and the inferred values for the BRM (with standard deviation $\sigma = 5$). The majority of all observations (98.4%) were recovered within ± 5 units.

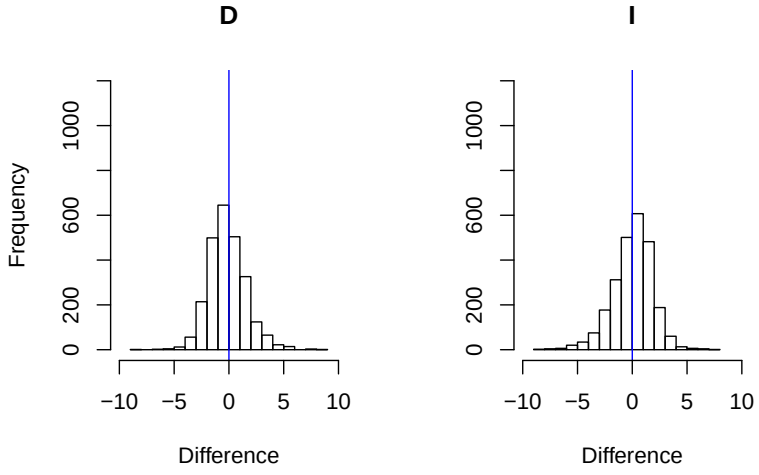

Figure 6.10: The difference between the recovered D and I values and the inferred values for the BRM (with standard deviation $\sigma = 10$). The majority of all observations (98.6%) were recovered within ± 5 units. The histogram of differences here appears to be approximating a normal distribution more closely than those displayed in Figures 6.9 and 6.8.

parameters are well recovered and that it is in only a relatively small number of cases that there was decreased predictive accuracy.

To ascertain where the skewed results in Figures 6.8 through 6.10 might be coming from, Tables 6.3 through 6.5 shows the average D, I and σ values recovered in the $\sigma = 1$, 5 and 10 cases for all combinations of D and I. There are two main patterns that seem to appear throughout these tables

1. As σ values increase, the recovery of all three parameters D, I, and σ becomes less accurate for values closer to zero and more accurate otherwise.

2. Parameter values appear most difficult to recover (especially in the cases where $\sigma = 1$ and to a lesser extent when $\sigma = 5$) when either $D = 0$ or $I = 0$ and the alternate parameter is at the extreme (such as in the $D = 0/I = 20$ or $D = -20/I = 0$ case).

Investigating these patterns further, Figures 6.11 through 6.13 consider the distribution of the differences between the values of D and I used to simulate the data and those that were recovered using the BRM. Confirming the patterns suggested across Figures 6.8 through 6.10 and in Tables 6.3 and 6.4, Figure 6.11 shows the majority of the box plots have an inter-quartile range reflective of the relatively small standard deviation ($\sigma = 1$) used when simulating the data. This is with the clear exception of when either $D = 0$ or $I = 0$. However, in those circumstances, the zero estimate (whether it be for $D = 0$ or $I = 0$) was well recovered with the increased variability observed in the alternate parameter. In addition, the alternate parameter was recovered with greater accuracy the closer it was to zero, suggesting that the hardest cases to recover were when D or I was most different from the corresponding $I = 0$ or $D = 0$ parameter.

As expected, corresponding to the increased σ value, the inter-quartile

Table 6.3: The average value of D (to one decimal place) recovered for pairs of D and I estimates on the BRM (with standard deviation $\sigma = 1$, 5 and 10). The average estimate of D is close to the imputed value in the majority of cases with the largest differences observed when either $D = 0$ or $I = 0$. When $I = 0$, D appears hardest to recover as $|D|$ becomes large but also becomes more accurate as σ increases.

Recovery of D

	I				
D	0	5	10	15	20
			$\sigma = 1$		
-20	-16.3	-19.3	-19.9	-19.9	-20.0
-15	-13.1	-14.9	-15.0	-15.0	-15.0
-10	-9.3	-10.0	-10.0	-10.0	-10.0
-5	-5.0	-5.0	-5.0	-5.0	-5.0
0	-0.0	-0.1	-0.2	-0.3	-0.4
			$\sigma = 5$		
-20	-16.1	-19.0	-19.7	-19.9	-20.0
-15	-13.0	-14.8	-14.9	-15.1	-15.0
-10	-9.4	-9.9	-10.0	-10.0	-10.0
-5	-5.0	-5.0	-5.1	-5.1	-5.0
0	-0.2	-0.4	-0.6	-0.7	-0.9
			$\sigma = 10$		
-20	-18.0	-19.3	-19.7	-19.9	-20.1
-15	-14.0	-4.5	-15.1	-15.0	-15.0
-10	-9.8	-10.0	-10.0	-10.1	-9.9
-5	-5.4	-5.3	-5.2	-5.3	-5.2
0	-0.9	-1.2	-1.4	-1.7	-1.8

Table 6.4: The average value of I (to one decimal place) recovered for pairs of D and I estimates on the BRM (with standard deviation $\sigma = 1$, 5 and 10). The average estimate of I is close to the imputed value in the majority of cases with the largest differences observed when either $D = 0$ or $I = 0$. When $D = 0$, I appears hardest to recover as I becomes large but also becomes more accurate as σ increases.

Recovery of I

	I				
D	0	5	10	15	20
			$\sigma = 1$		
-20	0.6	5.0	10.0	15.0	19.9
-15	0.4	5.0	10.0	15.0	19.9
-10	0.2	5.0	10.0	15.0	19.8
-5	0.1	5.0	10.0	14.8	19.3
0	0.1	4.9	8.9	11.8	14.4
			$\sigma = 5$		
-20	0.8	5.0	9.9	15.0	19.8
-15	0.7	5.1	9.8	15.1	19.8
-10	0.6	5.0	10.1	14.9	19.5
-5	0.4	5.1	10.0	14.6	18.9
0	0.3	5.0	9.2	12.5	15.4
			$\sigma = 10$		
-20	1.9	5.4	10.1	14.9	20.0
-15	1.5	5.3	10.4	14.9	19.8
-10	1.4	5.2	9.8	15.1	19.3
-5	1.1	5.2	10.1	14.4	18.3
0	1.1	5.5	9.4	13.3	17.2

Table 6.5: The average value of σ (to one decimal place) recovered for pairs of D and I estimates on the BRM (with standard deviation $\sigma = 1$, 5 and 10). The average estimate of σ is close to the imputed value in the majority of cases with the largest differences observed when either $D = 0$ or $I = 0$ and when $\sigma = 1$. The largest differences are specifically observed when σ is small and the imputed values for D and I are furthest from one another (for example $D = -20/I = 0$).

Recovery of σ					
	I				
D	0	5	10	15	20
	$\sigma = 1$				
-20	3.6	1.4	1.2	1.1	1.1
-15	2.6	1.1	1.0	1.0	1.1
-10	2.0	1.0	1.0	1.1	1.3
-5	1.7	1.0	1.0	1.1	1.5
0	1.5	1.2	1.9	2.1	2.9
	$\sigma = 5$				
-20	4.4	5.0	5.0	5.0	4.9
-15	4.4	4.9	4.8	5.0	4.9
-10	4.5	4.9	5.0	4.9	5.0
-5	4.7	4.9	5.0	4.9	5.0
0	4.8	4.7	4.5	4.3	4.4
	$\sigma = 10$				
-20	8.4	9.3	9.7	9.7	9.9
-15	8.4	9.5	9.8	9.8	9.8
-10	8.7	9.7	9.9	9.9	9.6
-5	8.9	9.8	9.6	9.5	9.4
0	9.5	9.2	8.7	8.5	8.3

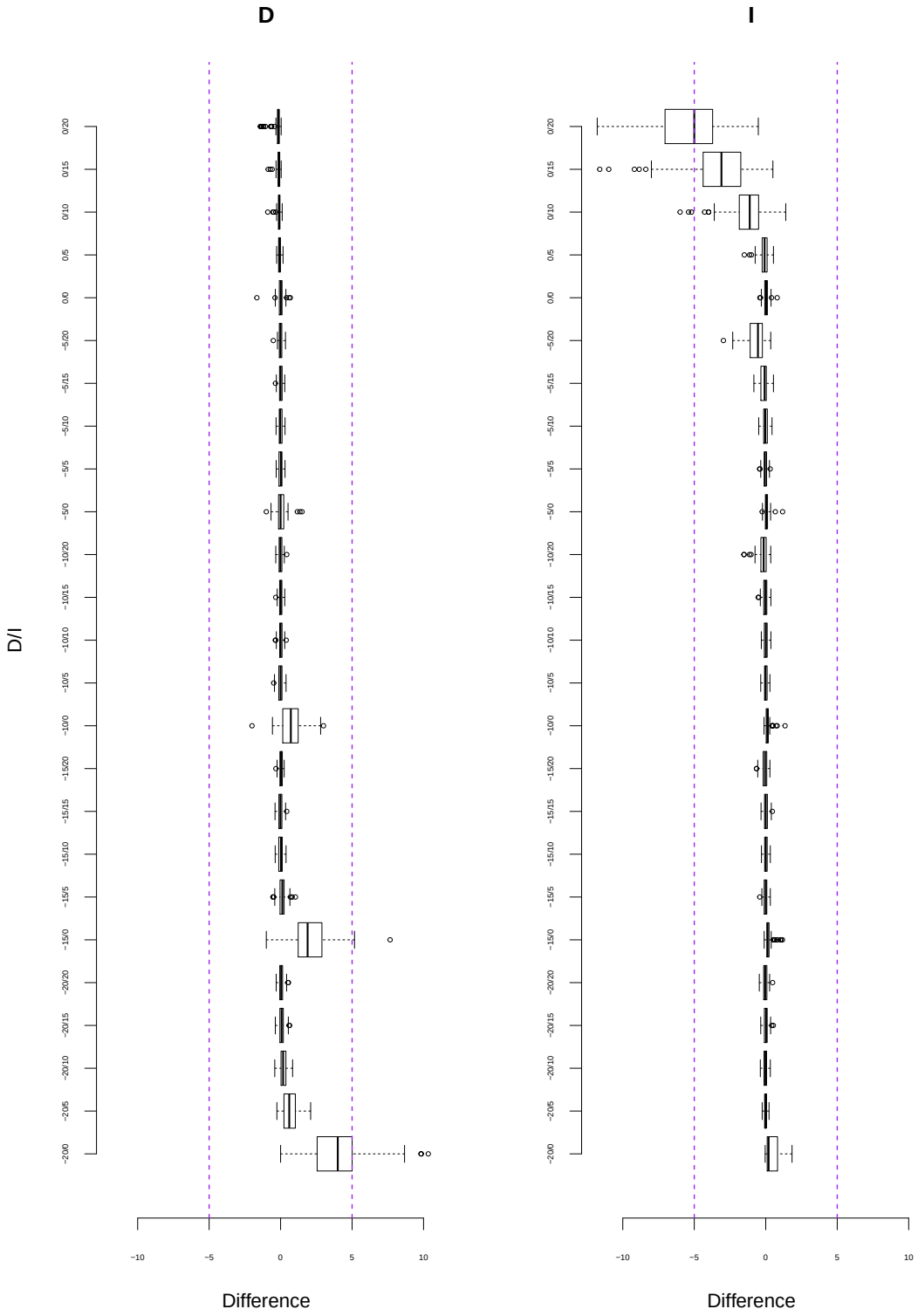

Figure 6.11: The difference between the recovered D (on the left) and I values and the imputed values for the BRM (with standard deviation $\sigma = 1$) for each combination of D and I. More variability is clearly observed when recovering the D parameter if $I = 0$ while more variability is observed when recovering the I parameter when $D = 0$. There is an observable pattern here where variability is largest when the difference between estimates is most extreme (for example $D = -20/I = 0$ and $D = 0/I = 20$).

range of the box-plots are wider in Figure 6.12 than those in Figure 6.11 but Figure 6.12 still shows that parameter recovery became more difficult when either the imputed values of D or I were equal to zero. Although less obvious in Figure 6.12 than in Figure 6.11, the same pattern of decreased accuracy in recovering the non-zero D or I parameter is matched to well recovered corresponding $I = 0$ or $D = 0$. However, when the standard deviation is increased to $\sigma = 10$, this pattern is much less visible (Figure 6.13).

Not only does Figure 6.13 lack the obvious pattern of increased variability associated with recovering parameters equal to zero observed in Figures 6.11 and 6.12, but it also lacks some of the more extreme individual difference scores. In both Figures 6.11 and 6.12 the inter-quartile range of values associated with $D = 0$ or $I = 0$ crosses over the line marking estimates as being more than 5 units from their imputed value. This is not the case in Figure 6.13 where only the tails of the distribution cross that line. This would seem to suggest that, with the increased variability across some parameter estimates, that parameter estimates are simultaneously becoming more localised in the extremes.

Whether the standard deviation used to simulate the data was $\sigma = 1$ (Figure 6.8), $\sigma = 5$ (Figures 6.9), or $\sigma = 10$ (Figure 6.10), the parameter values used to create the data were always recovered within a reasonable level of uncertainty. However it is obvious there are some problems recovering parameter values for low σ values when either $D = 0$ or $I = 0$ (Figures 6.11 and 6.12).

Run Dependent Response Model (RDRM)

As with the procedure for the 2pB and BRM, before analysing the results of the RDRM, if an individual run did not see at least five balloons burst

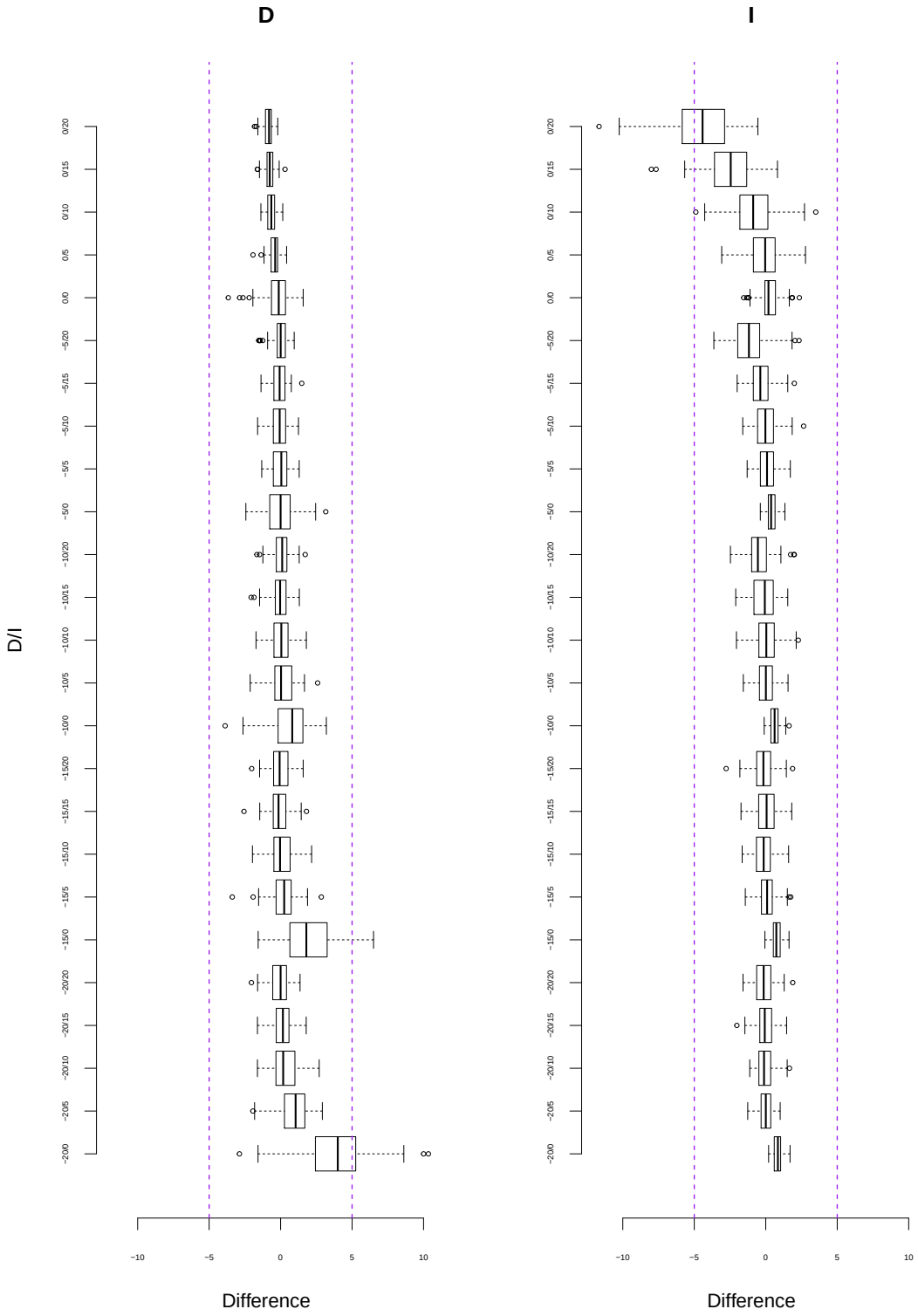

Figure 6.12: The difference between the recovered D (on the left) and I values and the imputed values for the BRM (with standard deviation $\sigma = 5$) for each combination of D and I. More variability is observed when recovering the D parameter if $I = 0$ while more variability is observed when recovering the I parameter when $D = 0$. As observed in Figure 6.11, when the estimates of D and I were most different from each other, recovery was most variable.

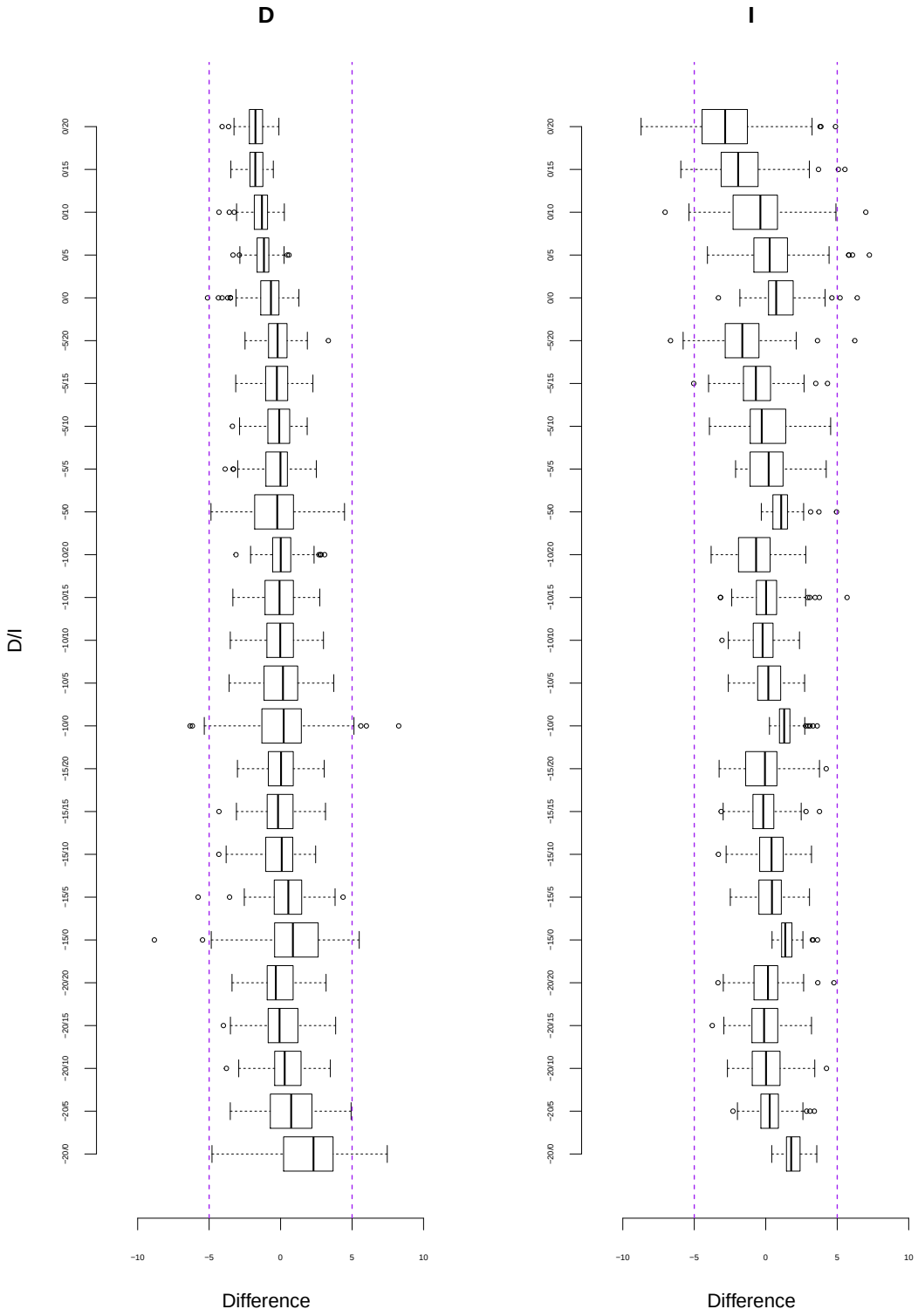

Figure 6.13: The difference between the recovered D (on the left) and I values and the imputed values for the BRM (with standard deviation $\sigma = 10$) for each combination of D and I. The difficulty observed in Figures 6.12 and 6.11 in estimating D when $I = 0$ or estimating I when $D = 0$ is no longer obvious here.

or cash in successfully at least five times, then the data was excluded from analyses. This process revealed 88.2% (5645/6400) cases as viable.

Figures 6.14 to 6.16 show the differences between the imputed parameters of the RDRM (using $\sigma = 1, 5$ or 10 respectively) and those recovered using maximum likelihood estimation. In general, the parameters D and I are well recovered within a reasonable band of error and, similar to Section 6.7.1, as σ increases, the histogram appears to approximate a normal distribution more closely.

Despite the accuracy associated with the D and I parameters, there are clear problems with the accuracy associated with recovering ϕ. Although the majority of estimates in Figures 6.14, 6.15 and 6.16 are close to zero, long tails in all three Figures suggest complications in some cases.

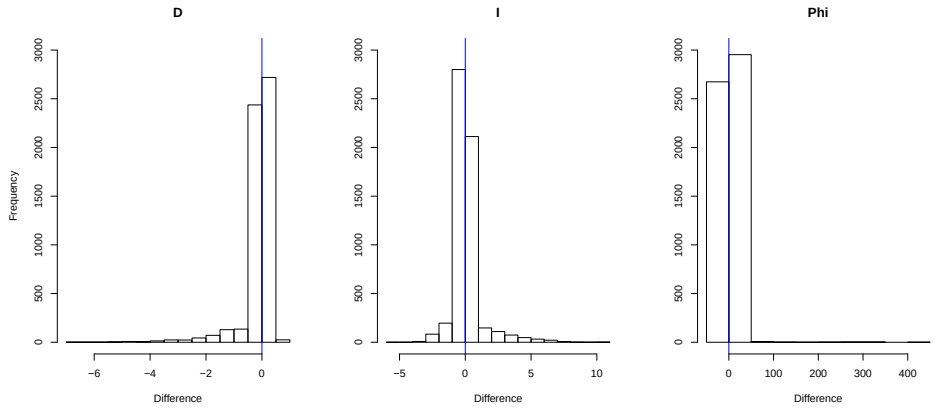

Figure 6.14: Histogram of the differences between the imputed parameter values of the RDRM with standard deviation $\sigma = 1$ and those recovered using maximum likelihood estimation. Although both D and I are recovered with good accuracy, there appears to be some cases in which large differences between the imputed values of ϕ and those recovered are observed.

Given that the relationship between the parameters D, I and ϕ is three dimensional, teasing out why the estimates of ϕ are more extreme than expected requires cosideration of the joint distribution of parameter estimates. Figures 6.17 through 6.25 show the distribution of the recovered parameters in pairwise fashion for all three implementations at $\sigma = 1, 5$ and $\sigma = 10$.

The pattern throughout all of the parameter recovery studies, that ϕ

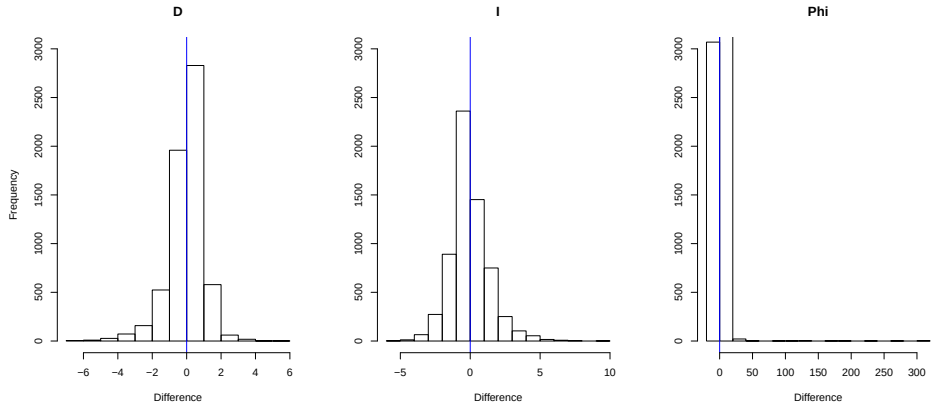

Figure 6.15: Histogram of the differences between the imputed parameter values of the RDRM with standard deviation $\sigma = 5$ and those recovered using maximum likelihood estimation. Although both D and I are recovered with good accuracy, there appears to be some cases in which large differences between the imputed values of ϕ and those recovered are observed.

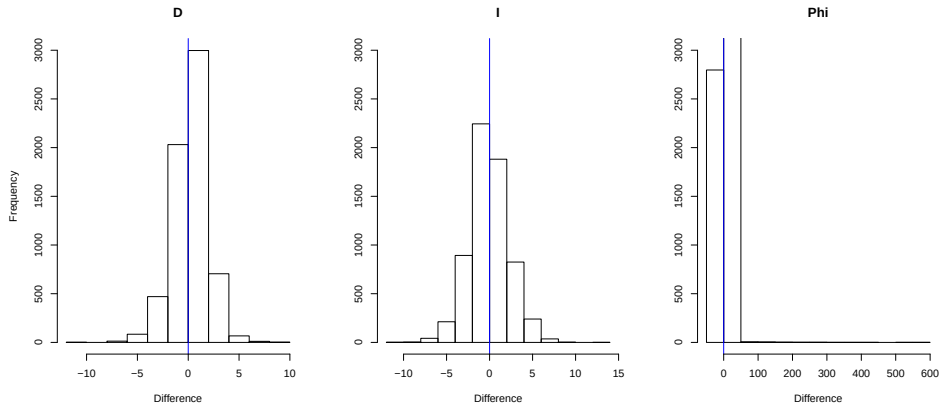

Figure 6.16: Histogram of the differences between the imputed parameter values of the RDRM with standard deviation $\sigma = 10$ and those recovered using maximum likelihood estimation. Although both D and I are recovered with good accuracy, there appears to be some cases in which large differences between the imputed values of ϕ and those recovered are observed.

was occasionally difficult to estimate, is confirmed in Figures 6.17 through 6.25. In addition however, as observed in Figure 6.19 for example, recovery of ϕ tended toward large, negative estimates.

When considering Equation 6.8:

$$\mu_{k+1} = Pump_k + \delta_k D + (1 - \delta_k)\, IR_k^{\phi},$$

the reason for these results may become more clear. Focusing on the IR_k^{ϕ}, it is clear that a large, negative estimate of ϕ would mean $R_k^{\phi} \to 0$ and, as

such, can be a proxy for an $I = 0$ estimate. The lack of identifiability in these cases is observable across all standard deviations implemented with large negative values of ϕ observable when $\sigma = 5$ and 10 in Figures 6.22 and 6.25 as well as in the $\sigma = 1$ case displayed in Figure 6.19.

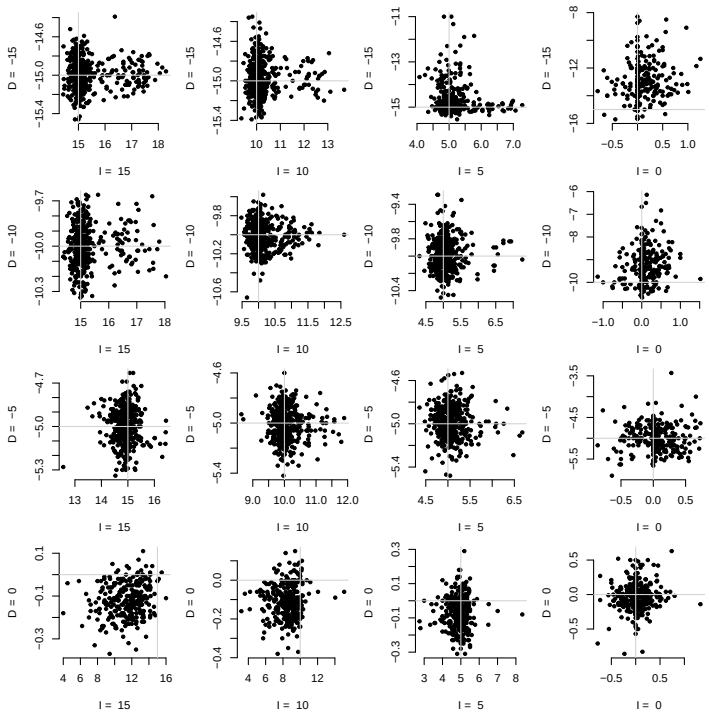

Figure 6.17: σ of 1, comparison between D and I for the RDRM with grey lines overlaid at the imputed values. The recovery of the parameters appears accurate with the most variability (obvious on the scale of the axes) in the $D = 0$ or $I = 0$ cases.

Apart from the lack of identifiability surrounding ϕ in some cases, Figures 6.17 through 6.25 confirm the general ability to recover parameters accurately that was observed in Figures 6.14 through 6.16. Most distributions of results are centred around the imputed values, especially in Figures 6.17, 6.20 and 6.23 when the joint distribution of D and I parameters are displayed.

6.7.2 Model Comparison

We have observed problems in estimating the ϕ parameter of the RDRM (Section 6.7.1), and given the similarity between the BRM and RDRM models, the following analysis looks at the model comparison between the 2pB

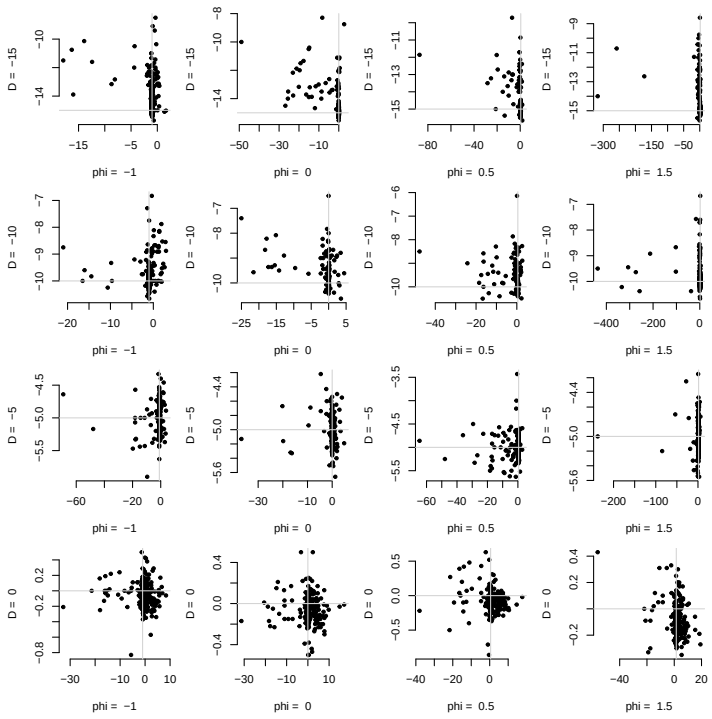

Figure 6.18: σ of 1, comparison between D and ϕ for the RDRM with grey lines overlaid at the imputed values. The recovery of parameters appears generally accurate with the exception of some large, negative estimates of ϕ in every comparison.

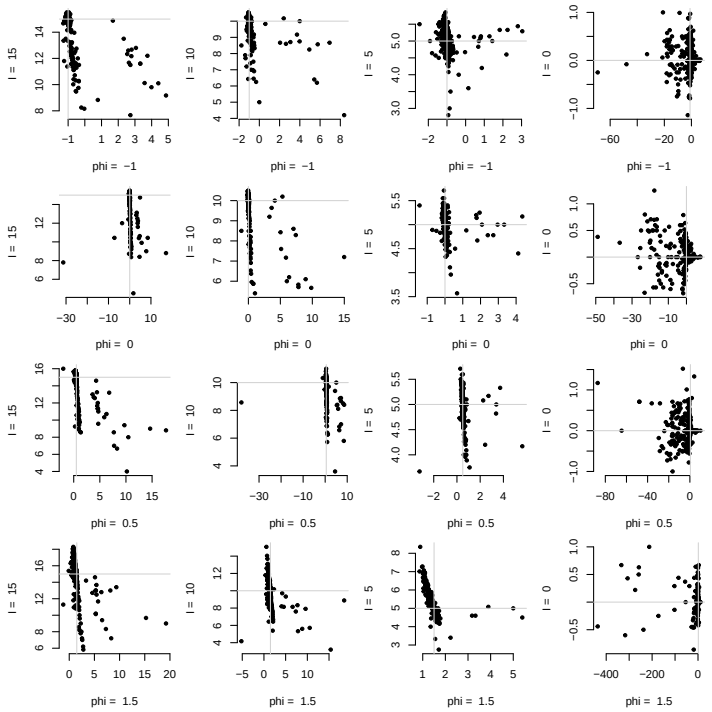

Figure 6.19: σ of 1, comparison between ϕ and I for the RDRM with grey lines overlaid at the imputed values. The recovery of parameters appears generally accurate with the exception of some large, negative estimates of ϕ when $I = 0$.

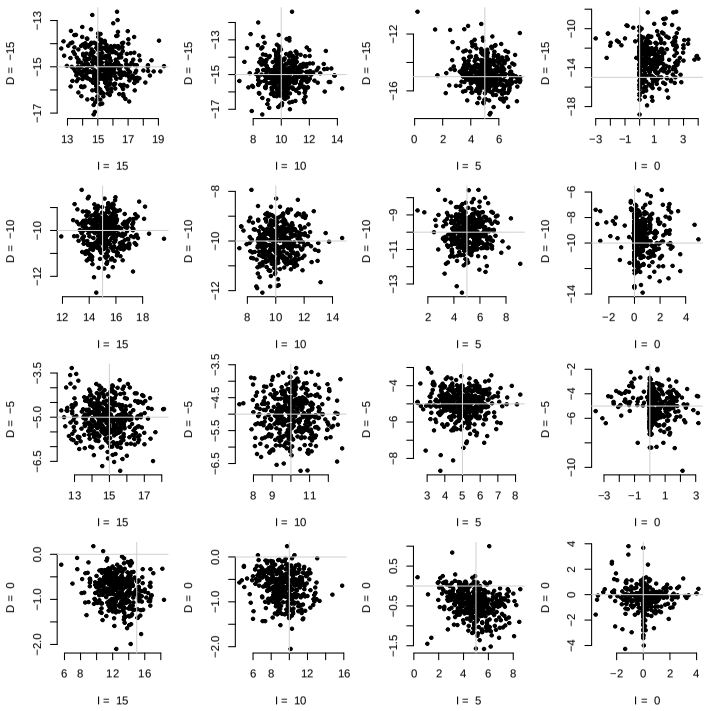

Figure 6.20: σ of 5, comparison between D and I for the RDRM with grey lines overlaid at the imputed values. The recovery of the parameters appears accurate with the most variability (obvious on the scale of the axes) in the $D = 0$ or $I = 0$ cases.

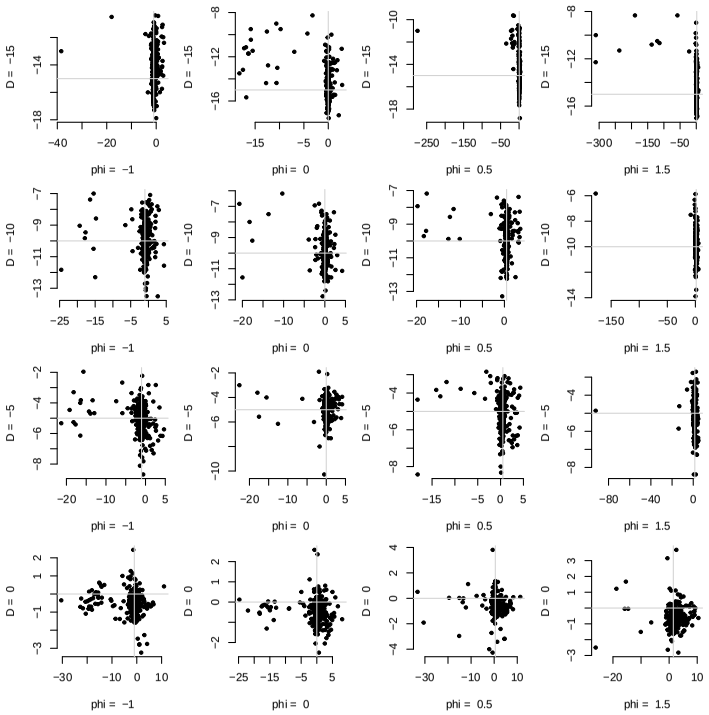

Figure 6.21: σ of 5, comparison between D and ϕ for the RDRM with grey lines overlaid at the imputed values. The recovery of parameters appears generally accurate with the exception of some large, negative estimates of ϕ in every comparison.

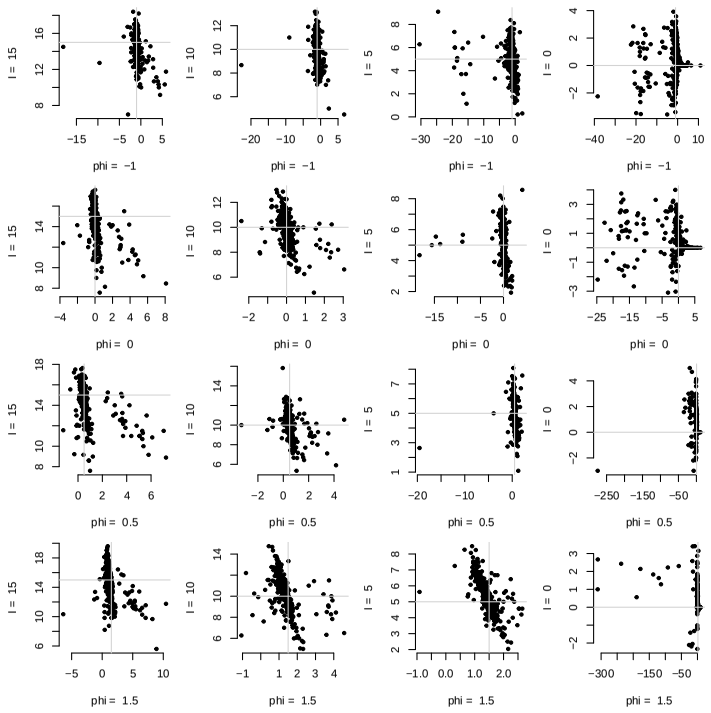

Figure 6.22: σ of 5, comparison between ϕ and I for the RDRM with grey lines overlaid at the imputed values. The recovery of parameters appears generally accurate with the exception of some large, negative estimates of ϕ when $I = 0$.

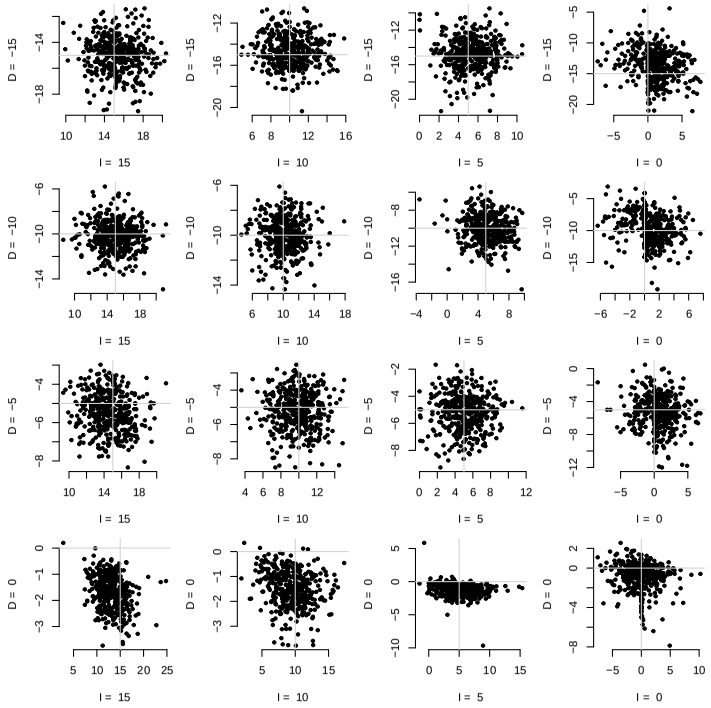

Figure 6.23: σ of 10, comparison between D and I for the RDRM with grey lines overlaid at the imputed values. The recovery of the parameters appears accurate.

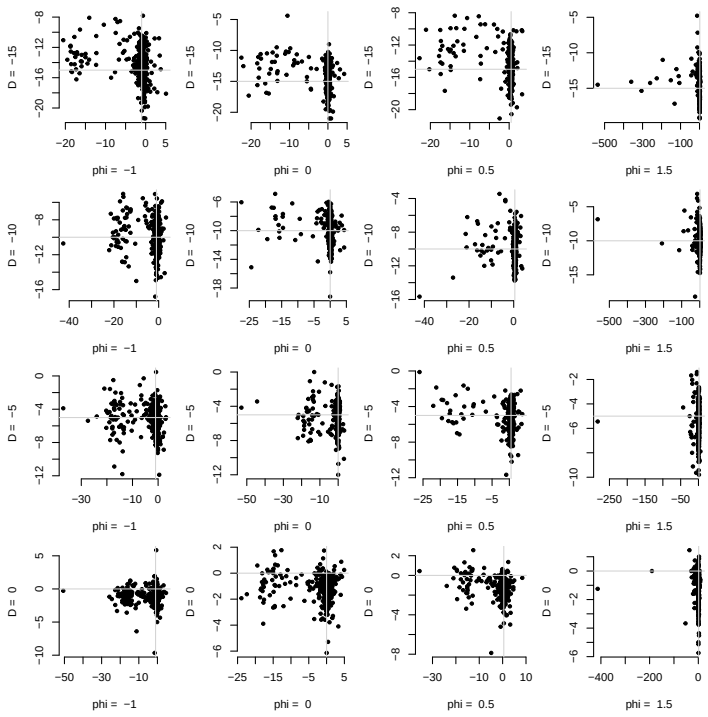

Figure 6.24: σ of 10, comparison between D and ϕ for the RDRM with grey lines overlaid at the imputed values. The recovery of parameters appears generally accurate with the exception of some large, negative estimates of ϕ.

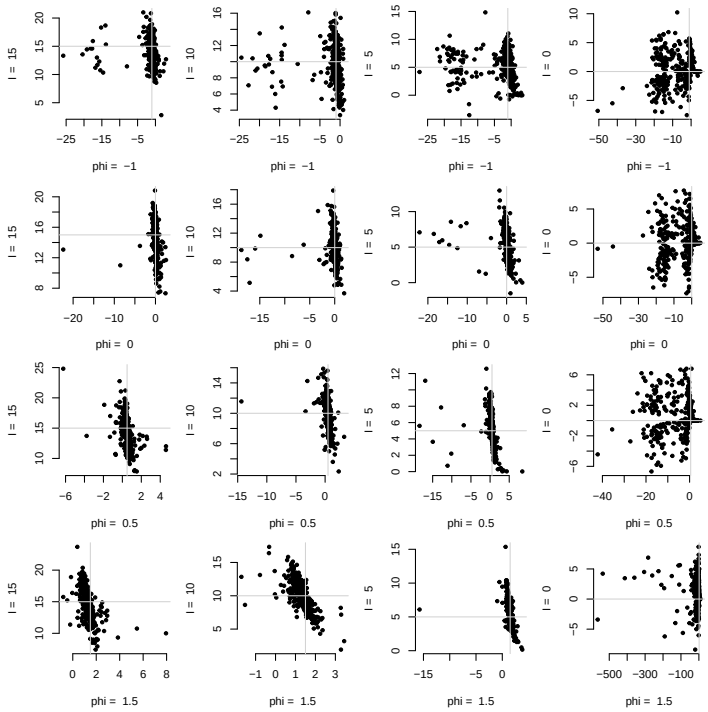

Figure 6.25: σ of 10, comparison between ϕ and I for the RDRM with grey lines overlaid at the imputed values. The recovery of parameters appears generally accurate with the exception of some large, negative estimates of ϕ when $I = 0$.

and the BRM.

Using the parameter combinations for the 2pB and BRM defined in Section 6.6, 20 data sets were simulated for each of the unique parameter combinations for both the 2pB (720 data sets in total) and the BRM (500 data sets in total). Both the 2pB and BRM were then fit to all of the data and their parameters estimates stored. This also provided the measurements for $\Delta GOF_{2pBvBRM}$ on both the 2pB and BRM data (AIC is used as the GOF measure). From here, the estimates were used to simulate twenty new data sets each (14,400 for the 2pB and 10,000 for the BRM) and then both models were fit back to all of the new data. This provided the estimates for the $\Delta GOF_{X_{2pB},2pBvBRM}$ and $\Delta GOF_{X_{BRM},2pBvBRM}$ for each of the original 2pB and BRM data sets.

Of the original 720 data sets simulated using the 2pB, 25 times the optimiser failed to converge to a solution when ascertaining the parameter estimates. This reduced the size of the original data set to 695 and, of the 14,400 data sets created using the 2pB from these, 482 times the optimiser failed to converge to a solution. This reduced the total number of data sets for analysis using the 2pB to 13,918. Similarly, although the optimiser converge for all 500 of the original data sets simulated using the BRM, the optimiser failed to converge when fitting the parameters of the 2pB to the data simulated using its own estimates 622 times. This reduced the total number of data sets from 10,000 to 9,378 in the set of data originally created using the BRM.

The ΔGOF_{AB} values for the fits to the original base BRM data showed a preference for the 2pB model over the BRM (median $(m) = -224$, inter-quartile range $= IQR[-396, -95]$) and when the base data was simulated using the 2pB ($m = -329$, $IQR[-468, -206]$). Table 6.6 shows that this preference for the 2pB continued when looking at the full the full $\Delta GOF_{X_A,AB}$

Table 6.6: Median and inter-quartile range (IQR) statistics for the difference in AIC ($\Delta GOF_{X_A,AB}$) for all model comparisons across both simulated data sets. The preference for the 2pB over the BRM is obvious from the negative values in the distributions.

	$\Delta GOF_{X_A,AB}$
2pB base data	
$\Delta GOF_{X_{2pB},2pBvBRM}$	$m = -60, [-114, 2]$
$\Delta GOF_{X_{BRM},2pBvBRM}$	$m = -160, [-228, -100]$
BRM base data	
$\Delta GOF_{X_{2pB},2pBvBRM}$	$m = -266, [-336, -194]$
$\Delta GOF_{X_{BRM},2pBvBRM}$	$m = -111, [-182, -48]$

distributions also.

However, even though the 2pB was preferred, the parameter estimates gained from the 2pB fits generally created probability functions that suggested random choices within the simulated data. Table 6.7 shows that the parameter estimates for α for the 2pB, whenever the BRM was used in any way to simulate the data, were often beyond the maximum number of allowable pumps in the BART (higher than 130). These high α were coupled with extremely low β estimates which result in a relatively flat probability of cashing in across all possible pumping values. A probability density of this type would manifest in extremely random choices of when to cash in the balloon. So, even though the 2pB may have been preferred (in terms of Goodness of fit) over the BRM, it is mainly because it is predicting a lack of pattern in the behaviours observed in the simulated data.

The estimates for the BRM also showed some unexpected unexpected results if the 2pB was used to simulate any of the data it was fitting to. When the base data was simulated using the 2pB, the parameter estimates for I in the BRM were extremely low and tightly distributed around zero. With a corresponding low median for D this would suggest that the majority of the pumping behaviour was being modelled by σ. This pattern seems to continue when the BRM simulated the base data but the second simulation was using the 2pB. Although the estimates for D and I are within the range

Table 6.7: Median and inter-quartile range ([IQR]) statistics for the parameter estimates for the 2pB and BRM across both data simulations. The base data used for the model comparisons were either 2pB or BRM and, from that data, the 2pB and BRM were fitted (Fitted Model) and a second set of data simulated using the fitting models. The parameter recovery was then completed for both the 2pB and BRM across all simulated data. It is clear that both models recover more accurately when the base data were created by the model that simulated them suggesting these models do not mimic each other well. When the base data were simulated using the 2pB, the BRM's I parameter gave the most unexpected results with small values centred around zero. When the base data were simulated using the BRM, the 2pB's α parameter inflated beyond the maximum number of pumps for the task and were coupled with very low β estimates suggesting flat probability functions and random participant choices.

			Simulated Using	
			2pB	BRM
Base	Fitted			
2pB	2pB	α	$m = 50, [10, 80]$	$m = 130, [71, 1258]$
		β	$m = 0.55, [0.24, 0.82]$	$m = 0.04, [0.00, 0.10]$
	BRM	D	$m = -1, [-12.3, 0.0]$	$m = -1, [-16.7, 0.0]$
		I	$m = 0.0, [-0.2, 0.3]$	$m = 0, [-0.1, 1.7]$
		σ	$m = 4.3, [2.5, 14.3]$	$m = 4.2, [2.4, 9.8]$
BRM	2pB	α	$m = 156, [133, 185]$	$m = 155, [137, 178]$
		β	$m = 0.04, [0.03, 0.05]$	$m = 0.04, [0.03, 0.05]$
	BRM	D	$m = -10, [-16.2, -5.5]$	$m = -10, [-15.4, -4.9]$
		I	$m = 9, [5.3, 14.8]$	$m = 10, [4.6, 15.5]$
		σ	$m = 32, [26.6, 36.1]$	$m = 9.1, [8.2, 9.8]$

of expected responses, the estimates for σ are much higher than expected suggesting that this may be what is explaining the difference between the two models.

These results, however, are averaged across all of the initial parameter combinations so to ascertain if these patterns were consistent across each of these combinations, the original GOF values were compared to the simulated distributions to ascertain which model is fitting the data best in which scenario. All original ΔGOF_{AB} values were classified against the $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ distributions that were generated from the 2pB and the BRM.

The ability to correctly classify the original GOF values as coming from one of the 2pB or BRM varied depending on how much overlap there was between the $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ distributions. Figure 6.26 shows an example two distributions with overlap but that are clearly distinct. In this instance the ability to classify data as being generated from one model or the other is possible for observations in the lower tail of the lower distribution or the upper tail of the upper distribution. The sort of overlapping-but-distinct distributions were most common when looking at the $\Delta GOF_{X_A,AB}$ when the base data were BRM allowing for a greater ability to classify observations than when the 2pB were used to simulated the original data.

There were cases when the distributions of $\Delta GOF_{X_A,AB}$ for the BRM and 2pB were almost identical. Figure 6.27 displays one such situation and it is clear, in this Figure, that there is very little difference between the distributions generated on the BRM data and the 2pB data. It is impossible, in this situation, to classify as BRM or 2pB and the result for all observations here would be that we can not determine which model is best. This sort of overlap occurred most frequently when the base data was simulated using the 2pB.

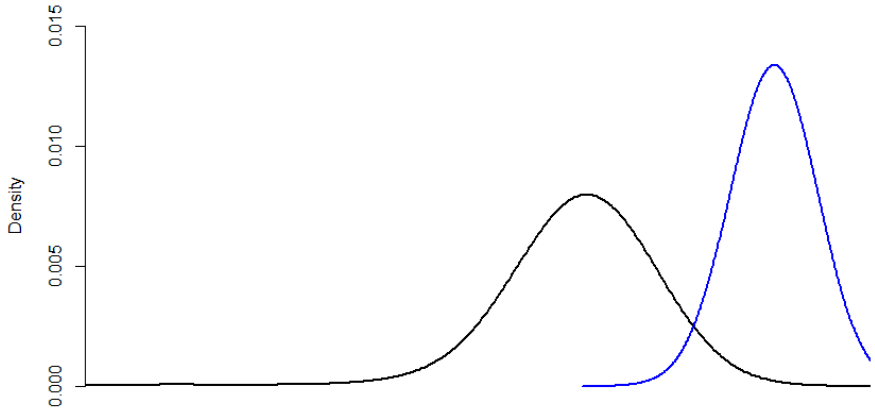

Figure 6.26: Two distributions with some overlap, but generally distinct, make for the best opportunity to be able to classify as either one or the other. In this example, smaller values would be classified as coming from a BRM model while larger values would be classified as 2pB.

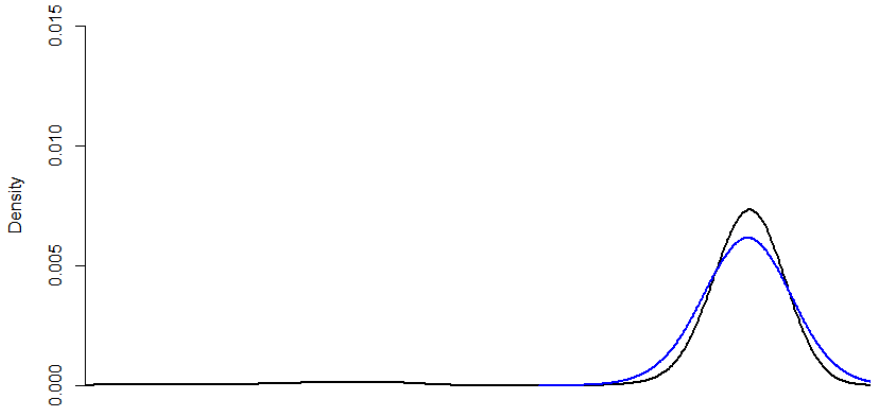

Figure 6.27: There is almost complete overlap between the distributions generated using the BRM 2pB models. In this situation it is impossible to classify as either model due to the close overlap.

The final situation encountered in classification of the ΔGOF_{AB} was associated with a $\Delta GOF_{X_A,AB}$ completely nested within the $\Delta GOF_{X_B,AB}$ distributions. Figure 6.28 gives an example of this where the distribution generated on the 2pB data is completely nested within the the distribution generated using the BRM. In this case, the only classifications that

can occur are when observations fall in the tails of the BRM distribution, leading to classification as BRM. Otherwise, observations here are unable to be identified. A result of having one distribution completely nested within another happened across both sets of base data but it was usually the BRM distribution that was the widest. This may be why the results which were averaged across all parameter combinations suggested that the 2pB was preferred over the BRM in general as the occasional large variability estimates in GOF for the BRM may skew the results.

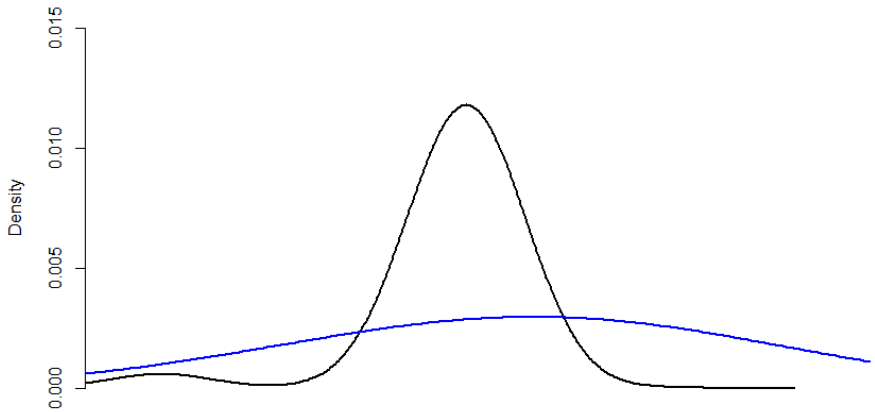

Figure 6.28: In the situation pictured here, the distribution generated on the 2pB data is completely nested within the the distribution generated using the BRM. In this case, the only classifications that can occur are either to identify as being from the BRM or be unidentifiable.

Although there was occasional nesting of the 2pB distribution inside the BRM distributions, the original GOF fit values between the 2pB and the BRM were classified the majority of the time (Table 6.8). In fact, Table 6.8 shows that, when the BRM simulated the base data only ten times was it unable to classify the original ΔGOF_{AB} and the BRM was chosen as the preferred model 79% of the time. When the 2pB simulated the base data there was only a slight preference for the 2pB over the BRM as the simulating model with 48% classified as coming from the 2pB distribution. However,

Table 6.8: The confusion matrix for the classification of base data given the goodness of fit model comparisons. Here the original ΔGOF_{AB} of the 2pB versus BRM, is classified using the related distributions of $\Delta GOF_{X_A,AB}$ and $\Delta GOF_{X_B,AB}$ for both the base data simulated using the 2pB and the BRM. When the base data was simulated using the 2pB, there was very little difference in preference for the 2pB or the BRM as the generating model. When the base data was simulated using the BRM, there was a clear preference for the BRM over the 2pB.

	Predicted Data		
Base Data	2pB (%)	BRM (%)	CND
2pB	333 (*48*)	318 (*46*)	44
BRM	93 (*19*)	397 (*79*)	10

it was the case that 91% (290 cases) of the observations classified as BRM when the base data was 2pB were from nested distributions such as the one displayed in Figure 6.28 while only this only occurred 3 times when the 2pB was preferred. This suggests that it was the large variation in the estimates associated with the BRM that influenced the incorrect categorisation with the 2pB data.

It would appear then, that data simulated using the 2pB or the BRM is best recovered when using the model that created it with some potential for the BRM to also explain 2pB data.

6.8 Discussion

The aim of this study was to ascertain if parameter estimates of any model of behaviour on the BART could be gained with enough accuracy to be useful at the level of the individual. This is very different to being able to use cognitive models of performance to measure differences between groups of individuals in a research context. Although van Ravenzwaaij et al. (2011) found success using the 2pB model to find differences between groups, the results of this study would suggest that this model will not always provide reliable results if used in a clinical context where the goal is to make inference about possible cognitive deficits. Importantly, it is often individuals with extreme behaviours that will present for clinical evaluation. For exam-

ple individuals with deficits in memory processes, impulsive behaviours or other cognitive deficits resulting in behavioural issues that, in turn, lead to psychological assessment and intervention. It is therefore imperative that a cognitive model of performance on a task be accurate at measuring parameters which reflect these extreme behaviours.

Section 6.7.1 showed that estimates of α and β in the 2pB model were not accurately recovered when parameter estimates became more extreme. As a reminder, in the 2pB, $\alpha \in (0, m)$ represents the point at, above which, the individual believes the balloon is more likely to burst than not and $\beta \in (0, 1)$ is a measure of the participant's behavioural consistency with that belief. If the maximum number of pumps m on any given trial is $m = 130$, then it would not be unreasonable to think a participant may have an α estimate of $40, 50$ or 60. Yet if we couple an estimate of α in this range with a high level of behavioural consistency (for example $\beta \geq 0.6$ it becomes increasingly difficult to recover β accurately with estimates 0.2 or greater different to the value used to simulate the data. A difference of 0.2 in an estimate of behavioural consistency is a large difference when $\beta \in (0, 1)$ and could lead to a clinical interpretation of these estimates that is not a true reflection of the individual's behaviour.

Similarly, when behavioural consistency was very low, Section 6.7.1 showed that it became difficult to recover α within a reasonable level of accuracy with estimates 20 points different to those used to simulate the data. Again, the interpretation of these estimates for an individual in a clinical setting may not reflect the individual's true cognitive process. In both the case of α and β, individuals with extreme estimates are not guaranteed to be recovered with a reasonable level of accuracy.

Using a simple empirical model, the BRM, to explain the data was most effective in parameter recover simulations. With parameter D empirically

modelling the decrease in number of pumps guessed following a burst and I the size of the increase following a successful cash-in, both parameters were recovered within ± 5 units in the majority of cases. The clear exception to this was when the standard deviation of the simulation was low and simulated values of D or I were $= 0$. In those situations, estimates were out by as much as 10 units difference for some specific combinations of parameters where either $D = 0$, or $I = 0$ and the corresponding I or D values were large.

When the σ parameter is low, an individual's pump guesses are highly consistent with what is predicted by the model, leading to guesses with a very small range. This can be a problem when one parameter is set to zero and the other is allowed to take large steps as the individual is more likely to converge to the maximal or minimal pump guesses allowed by the task (minimum of 1 pump or maximum of 130 pumps). For example, a data set generated with $D = 0$, $I = 20$ and $\sigma = 1$ will decrease their pump guess on trial $k + 1$ by $Change_{k+1} \sim N(0, 1)$ if trial k was a burst or alternately, if trial k was successfully cashed in, increase their pump guess by $Change_{k+1} \sim N(20, 1)$. Continuing this behaviour will potentially end up with guesses being capped at the maximal value of 130 making it harder for the model to estimate the true value of I. However, in those circumstances, the zero estimate (whether it be for $D = 0$ or $I = 0$) was well recovered allowing for some level of interpretation of cognitive deficits in the decision making process. Given the parameters of the BRM have been theorised to be linked to an individual's impulsivity following a loss (D) and confidence following a win (I), in the situation where $D = 0$ this would indicate an impulsive individual who is discounting losses, while an estimate of $I = 0$ would suggest the individual is discounting wins. As such, these estimates may still be of use in a clinical context suggesting that the BRM may be a

viable alternative to the 2pB for use at the level of the individual.

The RDRM was considered as an attempt to model trial dependency above what is modelled in the BRM by including information about how many successful trials (the run length) an individual has seen in a row. The inclusion of the ϕ parameter enabled the effect of the length of the run to have increasing, stable or decreasing effects on the magnitude of a pump guess following a win. However, the inclusion of the ϕ parameter resulted in some situations where the estimates of ϕ and I could not be uniquely identified. Similarly to the least accurate estimates using the BRM, ϕ was most difficult to recover when $I = 0$ with estimates occasionally substituting an $I = 0$ estimate with a large, negative estimate of ϕ. Despite this, the psychological interpretation of these estimates are arguably similar with both suggesting the individual is ignoring win information.

In Section 6.7.2 the BRM showed that it was more able to adapt to different styles of data than the 2pB. Although both models were best identified when they also simulated the data they were being fit to, the BRM was preferred when the original data was simulated using the BRM but there was no clear preference between the 2pB and the BRM when the original data was created using the 2pB. This finding was, however, tempered by the kinds of estimates gained through the parameter recovery simulations: 2pB estimates of BRM data suggested flat probability profiles with corresponding random participant choices, where as BRM fits to 2pB data tended to explain the variability in estimates through expansion of the σ parameter (a measure of variance). Neither of these situations are particularly insightful clinically. The next question is then, how do real people behave? Is behaviour too complex for the simplicity of the BRM to model effectively? Ideally, the next step in analysis should be to fit the BRM to some real data. The σ parameter can be estimated alongside D and I as a measure

of behavioural consistency and it can be investigated what range of values σ may be expected to take.

It would also be interesting to examine whether any of these models perform better if the task itself were more realistic. Balloons are familiar to many individuals and it is arguable that they do not burst in a uniformly distributed pattern. If the task itself were to be designed to have an optimal burst point with normally distributed errors, this would increase the complexity of the task but it would make it more realistic. The question then would be whether, like with the Iowa Gambling Task in Chapter 4, the task has become too complex to tease apart memory from impulsivity, however it would be worth exploring.

One of the main aims of the BRM model was to find an empirical way of modelling BART data and, even a small extension into a mix between mechanistic and empirical models withe RDRM seemed to make parameter recovery more complex. The maximum likelihood analysis reported here was chosen to facilitate this empirical style but it may be the case that moving to Bayesian analyses would provide stronger results. Of course, there are other frequentist styles of analyses that could be pursued also and, like the suggestion in Chapter 2, exploring a generalised additive model (GAM) to help define patterns in the estimates of σ (if they exist) for the BRM may provide the extra insight that will help extend this model. Using a GAM is a unique, empirical style of analysis which allows the definition of patterns or trends in the data based on the data itself. Either way, several questions remain unanswered that must be addressed before either the BRM or the RDRM can be proposed for clinical use. Firstly, a range of realistic values for σ needs to be ascertained such that the distribution assumptions of the models can be tailored to maximise the accuracy of results. Second, ascertaining whether the BRM or RDRM gains the most consistent and well

localised estimates when applied to real world data needs to be investigated. And, finally, a study of the validity of the measures needs to be conducted. Links between D, I and σ and the proposed cognitive processes which they represent (namely impulsivity or risk aversion, confidence and behavioural consistency) need to be established.

Chapter 7

Conclusion

The first part of this thesis was concerned with estimating the extent of illicit drug use in the community. Chapter 2 presented a published paper based on empirical, time series analysis of longitudinal illicit drug use data, obtained from wastewater analysis of a large catchment for four different drugs. We found that understanding weekly cycles in drug use can help inform the design of wastewater sampling schemes to maximise the ability to reliably interpret temporal trends while minimising the cost of data collection. For the first time, the analysis in Chapter 2 provides researchers with a comprehensive assessment of monitoring schemes for drugs with different usage patterns, helping researchers in the field to design long term monitoring of drug use in different sewage catchments in a cost-effective way.

Extending the sampling strategies suggested in Chapter 2 to new regions outside of the Gold Coast, Queensland, would be a logical next step for analysis. However, a more subtle extension on this topic is related to the kinds of information that can be obtained from the analysis itself. Time series provides a measure of the overall trend in the data and, while in Chapter 2 summaries of least squares results were constructed to provide measures of the differences between the sampling strategies tested, there was no concrete way of summarising the statistical significance of overall trends in drug use for each of the four drugs. In that way, time series provides a descriptive

summary of overall trend rather than an inferential one. The use of Generalised Additive Models (GAMs) in the place of time series analysis would allow for a summary of the overall trends in the data that can be used in a more inferential way (Hastie and Tibshirani, 1990). Highly adaptable, GAMs are only recently being discovered across the psychological sciences with a search for *Generalized additive model* in the PsychINFO data base returning only 28 scholarly journal results, 15 of which have been published since 2014. In the area of wastewater-based epidemiology, it appears that GAMs have not yet been investigated as a way of quantifying the trends in data. If future work could look at implementing both an effective sampling scheme *and* a more inferential form of analysis, then the information obtained from wastewater sampling may become even more useful for the field.

Moving to measuring the cognitive impact of illicit drug use on the individual, Chapters 4 - 6 used both empirical and mechanistic models to decompose behaviour on two popular psychological assessment tools which produce temporally autocorrelated data.

Chapter 4 presented a published paper addressing whether a mechanistic model of cognitive performance on the Iowa Gambling Task (IGT), the Expectancy Valence Model (EVM), is suitable for use at decomposing behaviour at the level of the individual. Parameter estimates of the EVM, when applied to IGT data, were revealed to be non-normal and unreasonably localised with observed estimates showing bi-modality, non-linearity or spanning the entire parameter space, even when an individual completed the IGT multiple times. As such, Chapter 4 found the EVM to be unsuitable for use in clinical practice where the goal is to base treatment decisions for an individual on the outcomes of their performance on the IGT. However, a new, simplified, two-parameter version of the EVM was suggested as a pos-

sible alternative to the EVM. The two-parameter EVM had clearer, more easily interpretable results when applied to IGT data suggesting it may be more viable for use in a clinical setting.

Chapter 4 also highlighted potential problems with the IGT itself. The seeming lack of test-retest reliability observed in experiment two and the lack of a clear relationship between frequency of deck choice and net return across all experiments were two of the most obvious indicators that the IGT may be a more complex task than it was designed to be. Despite this, the IGT is still in frequent use and many new cognitive models of performance are being proposed in the place of the EVM to tease apart the cognitive processes which give rise to behaviour on the IGT. These include the Prospect Valence Model (PVM) proposed by Ahn et al. (2008) and several variations of it. In future it would be interesting to see if the problems observed with recovering estimates at the level of the individual on the EVM are also obvious for the PVM and its related models. If the problems of non-normality and unreasonably localised estimates are prevalent across many models, it may be an indicator that the IGT itself need to be improved or simplified.

Chapter 6 continues the theme of simplification by using a simplified cognitive assessment tool, the Balloon Analogue Risk Task (BART) which relies on a less complex system of cognitive processes to produce behaviour than the IGT. Using simulation studies, Chapter 6 revealed that the variability associated with parameter estimates using the leading mechanistic model of cognitive performance on the BART, the van Ravenzwaaij et al. (2011) two-parameter BART model, led to a lack of accuracy in estimates of individual behaviour. As such, the van Ravenzwaaij et al. (2011) two-parameter BART model, may not always provide reliable results if used in a clinical context. The key phrase here being *in a clinical context.* The focus on parameter recovery for individuals who may behave more extremely than those who

are considered to be behaving within a *normal* range is important clinically as it often those with cognitive deficits who behave extremely which present for treatment. Accurately measuring those cognitive deficits are therefore extremely important if a model is to be used in a clinical context. Analysis of the two-parameter BART model revealed that, as the parameter values used for generating the data became more extreme, the accuracy of their recovery decreased. However two, new, empirically derived models of performance on the BART, the Basic Response Model (BRM) and the Run Dependent Response Model (RDRM), were able to recover parameter values with little variability under most conditions except for some extreme parameter values. Investigation is required to ascertain if those parameter values are likely to be observed in a clinical context.

The BMR had the least variable parameter recovery, especially when the standard deviation during simulation was not restricted too extremely. But, even in the case of a narrow standard deviation, the results gained still had an accurately interpretable outcome. The RDRM was less accurate at recovering parameters than the BRM but, in specific cases where identifiability of parameters became an issue, the physical interpretation of the results was consistent. These results suggest that both the BRM and RDRM may be possible alternatives for use in a clinical context over the two-parameter BART model. However, as both the BRM and RDRM are empirical models, and therefore were not constructed through consideration of the theories of the underling cognitive processes which gave rise to the observed data, Chapter 6 highlights the need for validity studies and application of the BRM and RDRM to real data as future endeavours.

In the introduction to this thesis, a discussion between Ferguson (2015) and Tryon (2016) raised a question about the future of psychological science; whether more mechanistic or more empirical models should be used. This

thesis has presented a situation in which empirical analysis has answered all of the questions asked in Chapter 2 (measuring the extent of illicit drug use in the community) and a situation in which a mechanistic model was proposed as a viable option in Chapter 4 (using the two-parameter EVM to gain more accurate estimates of performance on the IGT at the level of the individual). It is clear that there are some situations in which empirical analyses are required and provide the best answers to the research questions being asked but it is also clear that, to understand the cognitive mechanisms which give rise to observed behaviour, mechanistic models have a power that empirical analyses do not. But Chapter 6 presents an example in which a mechanistic model may not be able to be used with a good level of accuracy in a clinical context and two empirical alternatives that appear to be more accurate but which only have suggested links to the cognitive aspects of performance they might represent. So it would appear that the answer to the 'mechanistic or empirical' question is not straight forward.

Mechanistic models come from a perspective of hard theory, there are reasoned arguments behind the components included in the model, and these arguments and theories require testing and revision if they are found to be incorrect. The empirical, data driven models work by providing a solution to explaining observed data but with the added complexity of then having to determine what the results mean for the proposed theories. Ferguson (2015) argued for a need for flexibility in psychological sciences, with less reliance on existing, mechanistic models, to find evidence and Tryon (2016) responded with seemingly contrary suggestion that there is a need for more mechanistic models to compete with those existing in the psychological literature. But really mechanistic and empirical models are just direct or indirect ways of approaching the same question. As an example, the BRM and RDRM seem to answer both of these arguments. Although empirically derived,

hence answering Ferguson's request for less reliance on existing mechanistic models, the BRM and RDRM do appear to have the potential to be linked to the cognitive processes underlying behaviour. If those links can be validated then, satisfying Tryon's call for more mechanistic models, the BRM and/or RDRM may become new mechanistic models of cognitive performance on the BART.

Bibliography

Ahn, W. Y., Busemeyer, J., Wagenmakers, E. J., and Stout, J. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science: A Multidisciplinary Journal*, 32(8):1376–1402.

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., and Brown, J. W. (2011). A model-based fMRI analysis with hierarchical bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, 4(2):95–110.

Ansell, E. B., Laws, H. B., Roche, M. J., and Sinha, R. (2015). Effects of marijuana use on impulsivity and hostility in daily life. *Drug and Alcohol Dependence*, 148:136 – 142.

Balaguero, M. A., Vicente, M. J., Molina, A. G., Tormos, J. M., and Rovira, T. R. (2016). Balloon analogue risk task to assess decision-making in acquired brain injury. *International Journal of Psychological Research*, 9:30 – 39.

Baldacchino, A., Balfour, D. J. K., and Matthews, K. (2015). Impulsivity and opioid drugs: Differential effects of heroin, methadone and prescribed analgesic medication. *Psychological Medicine*, 45(6):1167–1179.

Banta-Green, C. J., Field, J. A., Chiaia, A. C., Sudakin, D. L., Power, L., and De Montigny, L. (2009). The spatial epidemiology of cocaine, methamphetamine and 3,4-methylenedioxymethamphetamine (MDMA)

use: A demonstration using a population measure of community drug load derived from municipal wastewater. *Addiction*, 104(11):1874–1880.

Bechara, A. (2012). *Iowa Gambling Task*^{TM}. PAR. [Software]. Available from http://www4.parinc.com/Products/Producr.aspx?ProductID=IGT.

Bechara, A., Damasio, A. R., Damasio, H., and Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3):7–15. PMID: 8039375.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, 10(3):295 –307.

Been, F., Bijlsma, L., Benaglia, L., Berset, J.-D., Botero-Coy, A. M., Castiglioni, S., Kraus, L., Zobel, F., Schaub, M. P., Bücheli, A., Hernández, F., Delémont, O., Esseiva, P., and Ort, C. (2016). Assessing geographical differences in illicit drug consumption: A comparison of results from epidemiological and wastewater data in germany and switzerland. *Drug and Alcohol Dependence*, 161:189 – 199.

Belin, D., Mar, A. C., Dalley, J. W., Robbins, T. W., and Everitt, B. J. (2008). High impulsivity predicts the switch to compulsive cocaine-taking. *Science*, 320(5881):1352–1355.

Bijlsma, L., Botero-Coy, A. M., Rincón, R. J., Peñuela, G. A., and Hernández, F. (2016). Estimation of illicit drug use in the main cities of colombia by means of urban wastewater analysis. *Science of The Total Environment*, 565:984 – 993.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brown, W. S., Anderson, L. B., Symington, M. F., and Paul, L. K. (2012). Decision-making in individuals with agenesis of the corpus callosum: Expectancy-valence in the Iowa gambling task. *Archives of Clinical Neuropsychology*, 27(5):532–544.

Buelow, M. T. and Suhr, J. A. (2009). Construct validity of the Iowa gambling task. *Neuropsychology review*, 19(1):102–114. PMID: 19194801.

Busemeyer, J. R. and Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, 14(3):253–262. PMID: 12214432.

Capone, F., Capone, G., Di Pino, G., Florio, L., Oricchio, G., and Di Lazzaro, V. (2016). Linking cognitive abilities with the propensity for risk-taking: The balloon analogue risk task. *Neurological Sciences*, pages 1–5.

Castiglioni, S., Zuccato, E., Crisci, E., Chiabrando, C., Fanelli, R., and Bagnati, R. (2006). Identification and measurement of illicit drugs and their metabolites in urban wastewater by liquid chromatography/tandem mass spectrometry. *Analytical Chemistry*, 78(24):8421–8429.

Chen, C., Kostakis, C., Harpas, P., Felgate, P. D., Irvine, R. J., and White, J. M. (2011). Marked decline in 3,4-methylenedioxymethamphetamine (MDMA) based on wastewater analysis. *Journal of Studies on Alcohol and Drugs*, 72(5):737–740.

Chiaia, A., Banta-Green, C., and Field, J. (2008). Eliminating solid phase extraction with large volume injection lc/msms: Analysis of illicit and legal drugs and human urine indicators in us wastewaters. *Environmental Science and Technology*, 42(23):8841–8848.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):pp. 327–335.

Chiu, Y.-C. and Lin, C.-H. (2007). Is deck C an advantageous deck in the Iowa gambling task? *Behavioural and Brain Functions*, 3(1):1–37.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Irma, T. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):1362–1368.

Congdon, P. (2003). *Applied Bayesian Modelling*. Wiley, 1 edition.

Cowpertwait, P. S. P. and Metcalfe, A. V. (2009). *Introductory Time Series with R*. Springer, New York.

Dai, J., Kerestes, R., Upton, D. J., Busemeyer, J. R., and Stout, J. C. (2015). An improved cognitive model of the iowa and soochow gambling tasks with regard to model fitting performance and tests of parameter consistency. *Frontiers in Psychology*, 6:229.

Dawe, S. and Loxton, N. J. (2004). The role of impulsivity in the development of substance use and eating disorders. *Neuroscience & Biobehavioral Reviews*, 28(3):343 – 351.

De Wilde, B., Bechara, A., Sabbe, B., Hulstijn, W., and Dom, G. (2013a). Risky decision-making but not delay discounting improves during inpatient treatment of polysubstance dependent alcoholics. *Frontiers in psychiatry*, 4:91. PMID: 24027538 PMCID: PMC3760441.

De Wilde, B., Verdejo-García, A., Sabbe, B., Hulstijn, W., and Dom, G. (2013b). Affective decision-making is predictive of three-month relapse in polysubstance-dependent alcoholics. *European addiction research*, 19(1):21–28. PMID: 22948315.

de Wit, H. (2009). Impulsivity as a determinant and consequence of drug use: a review of underlying processes. *Addiction Biology*, 14(1):22–31.

Economidou, D., Pelloux, Y., Robbins, T. W., Dalley, J. W., and Everitt, B. J. (2009). High impulsivity predicts relapse to cocaine-seeking after punishment-induced abstinence. *Biological Psychiatry*, 65(10):851 – 856.

Evenden, J. (1999a). Impulsivity: A discussion of clinical and experimental findings. *Journal of psychopharmacology (Oxford, England)*, 13(2):180–192. PMID: 10475725.

Evenden, J. L. (1999b). Varieties of impulsivity. *Psychopharmacology*, 146(4):348–361. PMID: 10550486.

Ferguson, C. J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist*, 70(6):527–542.

Fridberg, D. J., Queller, S., Ahn, W. Y., Kim, W., Bishara, A. J., Busemeyer, J. R., Porrino, L., and Stout, J. C. (2010). Cognitive mechanisms underlying risky decision-making in chronic cannabis users. *Journal of Mathematical Psychology*, 54(1):28–38.

Frost, N., Griffiths, P., and Fanelli, R. (2008). Peering into dirty waters: the potential and implications of a new approach to monitoring drug consumption. *Addiction*, 103(8):1239–1241.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman and Hall/CRC, 1 edition.

Goudriaan, A. E., Grekin, E. R., and Sher, K. J. (2011). Decision making and response inhibition as predictors of heavy alcohol use: A prospective

study. *Alcoholism, clinical and experimental research*, 35(6):1050–1057. PMID: 21332527 PMCID: PMC3097267.

Grasman, R. P. P. P. and Wagenmakers, E. J. (2005). A DHTML implementation of the Iowa gambling task.

Hall, W., Prichard, J., Kirkbride, P., Bruno, R., Thai, P. K., Gartner, C., Lai, F. Y., Ort, C., and Mueller, J. F. (2012). An analysis of ethical issues in using wastewater analysis to monitor illicit drug use. *Addiction*, 107(10):17671773.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

He, Q., Xue, G., Chen, C., Lu, Z.-L., Chen, C., Lei, X., Liu, Y., Li, J., Zhu, B., Moyzis, R. K., Dong, Q., and Bechara, A. (2012). COMT Val158Met polymorphism interacts with stressful life events and parental warmth to influence decision making. *Scientific Reports*, 2.

Humphries, M. A., Bruno, R., Karpievitch, Y., and Wotherspoon, S. (2015). The expectancy valence model of the Iowa gambling task: Can it produce reliable estimates for individuals? *Journal of Mathematical Psychology*, 64 65:17 – 34.

Humphries, M. A., Bruno, R., Lai, F. Y., Thai, P. K., Holland, B. R., O'Brien, J. W., Ort, C., and Mueller, J. F. (2016). Evaluation of monitoring schemes for wastewater-based epidemiology to identify drug use trends using cocaine, methamphetamine, MDMA and methadone. *Environmental Science & Technology*, 50(9):4760–4768.

Hunt, M. K., Hopko, D. R., Bare, R., Lejuez, C. W., and Robinson, E. V.

(2005). Cconstruct validity of the balloon analog risk task (BART): Associations with psychopathy and impulsivity. *Assessment*, 12:416–428.

Kaplan, R. M. and Saccuzzo, D. P. (2012). *Psychological Testing: Principles, Applications, and Issues*. Cengage Learning, Belmont, CA, 8 edition edition.

Kasar, M., Gleichgerrcht, E., Keskinkilic, C., Tabo, A., and Manes, F. F. (2010). Decision-making in people who relapsed to driving under the influence of alcohol. *Alcoholism, clinical and experimental research*, 34(12):2162–2168. PMID: 21087291.

Kessler, L., Hewig, J., Weichold, K., Silbereisen, R. K., and Miltner, W. H. R. (2016). Feedback negativity and decision-making behavior in the balloon analogue risk task (BART) in adolescents is modulated by peer presence. *Psychophysiology*.

Khan, U., van Nuijs, A. L., Li, J., Maho, W., Du, P., Li, K., Hou, L., Zhang, J., Meng, X., Li, X., and Covaci, A. (2014). Application of a sewage based approach to assess the use of ten illicit drugs in four chinese megacities. *Science of the Total Environment*, 487:710–721.

Kim, K. Y., Lai, F. Y., Kim, H.-Y., Thai, P. K., Mueller, J. F., and Oh, J.-E. (2015). The first application of wastewater-based drug epidemiology in five south korean cities. *Science of the Total Environment*, 524-525:440–446.

Kim, M.-S., Kang, B.-N., and Lim, J. Y. (2016). Decision-making deficits in patients with chronic schizophrenia: Iowa gambling task and prospect valence learning model. *Neuropsychiatric Disease and Treatment*, 12:1019–1027.

Kluwe-Schiavon, B., Viola, T. W., Sanvicente-Vieira, B., Pezzi, J. C., and

Grassi-Oliveira, R. (2016). Similarities between adult female crack cocaine users and adolescents in risky decision-making scenarios. *Journal of Clinical and Experimental Neuropsychology*, 38(7):795–810.

Lai, F. L., O'Brien, J. Thai, P. K., Hall, W., Chan, G., Bruno, R., Ort, C., Prichard, J., Carter, S., Anuj, S., Kirkbride, P., Gartner, C., Humphries, M., and Mueller, J. F. (2016a). Cocaine, MDMA and methamphetamine residues in wastewater: Consumption trends (2009–2015) in south east queensland, australia. *Science of the Total Environment*, 568:803–809.

Lai, F. L., O'Brien, J., Bruno, R., Hall, W., Prichard, J., Kirkbride, P., Gartner, C., Thai, P., Carter, S., Lloyd, B., Burns, L., and Mueller, J. (2016b). Spatial variations in the consumption of illicit stimulant drugs across australia: A nationwide application of wastewater-based epidemiology. *Science of The Total Environment*, 568:810 – 818.

Lai, F. Y., Anuj, S., Bruno, R., Carter, S., Gartner, C., Hall, W., Kirkbride, K. P., Mueller, J. F., O'Brien, J. W., Prichard, J., Thai, P. K., and Ort, C. (2015). Systematic and day-to-day effects of chemical-derived population estimates on wastewater-based drug epidemiology. *Environmental Science and Technology*, 49:999–1008.

Lai, F. Y., Ort, C., Gartner, C., Carter, S., Prichard, J., Kirkbride, P., Bruno, R., Hall, W., Eaglesham, G., and Mueller, J. F. (2011). Refining the estimation of illicit drug consumptions from wastewater analysis: Coanalysis of prescription pharmaceuticals and uncertainty assessment. *Water Research*, 45(15):4437–4448.

Lamers, C. T. J., Bechara, A., Rizzo, M., and Ramaekers, J. G. (2006). Cognitive function and mood in MDMA/THC users, THC users and non-

drug using controls. *Journal of Psychopharmacology (Oxford, England)*, 20(2):302–311. PMID: 16510488.

Lane, S. D., Yechiam, E., and Busemeyer, J. R. (2006). Application of a computational decision model to examine acute drug effects on human risk taking. *Experimental and Clinical Psychopharmacology*, 14:254–264.

Lauriola, M., Panno, A., Levin, I. P., and Lejuez, C. W. (2013). Individual differences in decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. *Journal of Behavioural Decision Making*, 27:20–36.

Laws, K. R. and Kokkalis, J. (2007). Ecstasy (MDMA) and memory function: a meta-analytic update. *Human Psychopharmacology: Clinical and Experimental*, 22(6):381–388.

Lehmann, E. L. (1990). Model specification: The views of fisher and neyman, and later developments. *Statistical Science*, 5(2):160–168.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D., and Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The baloon analogue risk tast (BART). *Journel of Experimental Psycology:Applied*, 8(2):75–84.

Lin, C.-H., Song, T.-J., Chen, Y.-Y., Lee, W.-K., and Chiu, Y. (2013). Re-examining the validity and reliability of the clinical version of the Iowa gambling task: Evidence from a normal subject group. *Frontiers in Psychology*, 4(220).

McCall, A.-K., Bade, R., Kinyua, J., Lai, F. Y., Thai, P. K., Covaci, A., Bijlsma, L., van Nuijs, A. L. N., and Ort, C. (2016). Critical review on the stability of illicit drugs in sewers and wastewater samples. *Water Research*, 88:933–947.

Moeller, F. G., Dougherty, D. M., Barratt, E. S., Schmitz, J. M., Swann, A. C., and Grabowski, J. (2001). The impact of impulsivity on cocaine use and retention in treatment. *Journal of Substance Abuse Treatment*, 21(4):193 – 198.

Nejtek, V. A., Kaiser, K. A., Zhang, B., and Djokovic, M. (2013). Iowa gambling task scores predict future drug use in bipolar disorder outpatients with stimulant dependence. *Psychiatry research*, 210(3):871–879. PMID: 24012163.

Nuijten, M., Blanken, P., Van den Brink, W., Goudriaan, A. E., and Hendriks, V. M. (2016). Impulsivity and attentional bias as predictors of modafinil treatment outcome for retention and drug use in crack-cocaine dependent patients: Results of a randomised controlled trial. *Journal of Psychopharmacology*, 30(7):616–626.

Ort, C., Eppler, J. M., Scheidegger, A., Rieckermann, J., Kinzig, M., and Sörgel, F. (2014a). Challenges of surveying wastewater drug loads of small populations and generalizable aspects on optimizing monitoring design. *Addiction*, 109(3):472–481.

Ort, C., van Nuijs, A. L. N., Berset, J.-D., Bijlsma, L., Castiglioni, S., Covaci, A., de Voogt, P., Emke, E., Fatta-Kassinos, D., Griffiths, P., Hernández, F., González-Mariño, I., Grabic, R., Kasprzyk-Hordern, B., Mastroianni, N., Meierjohann, A., Nefau, T., Östman, M., Pico, Y., Racamonde, I., Reid, M., Slobodnik, J., Terzic, S., Thomaidis, N., and Thomas, K. V. (2014b). Spatial differences and temporal changes in illicit drug use in europe quantified by wastewater analysis. *Addiction*, 109(8):1338–1352.

Parrott, A. C., Milani, R. M., Gouzoulis-Mayfrank, E., and Daumann, J. (2007). Cannabis and ecstasy/MDMA (3,4-

methylenedioxymethamphetamine): An analysis of their neuropsychobiological interactions in recreational users. *Journal of Neural Transmission*, 114(8):959–968. PMID: 17520319.

Passetti, F., Clark, L., Mehta, M. A., Joyce, E., and King, M. (2008). Neuropsychological predictors of clinical outcome in opiate addiction. *Drug and Alcohol Dependence*, 94(1-3):82–91. PMID: 18063322.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, USA, 1 edition.

Peacock, A., Bruno, R., Martin, F., and Carr, A. (2013). The impact of alcohol and energy drink consumption on intoxication and risk-taking behavior. *Alcoholism*, 37:1234–1242.

Peatfield, N., Parkinson, J., and Intriligator, J. (2012). Emotion-based learning is biased by brand logos. *Applied Cognitive Psychology*, 26(5):694–701.

Pleskac, T. J., Wallsten, T. S., Wang, P., and Lejuez, C. W. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, 16(6):555–564.

Poletti, M., Cavedini, P., and Bonuccelli, U. (2011). Iowa gambling task in Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology*, 33(4):395–409. PMID: 21140314.

Prichard, J., Lai, F. Y., Kirkbride, P., Bruno, R., Ort, C., Carter, S., Hall, W., Gartner, C., Thai, P. K., and Mueller, J. F. (2012). Measuring drug use patterns in queensland through wastewater analysis. *Trends and Issues in Crime and Criminal Justice*, 442:1–8.

Quednow, B. B., Kähn, K. U., Hoppe, C., Westheide, J., Maier, W., Daum, I., and Wagner, M. (2007). Elevated impulsivity and impaired decision-

making cognition in heavy users of MDMA ("Ecstasy"). *Psychopharmacology*, 189(4):517–530. PMID: 16425060.

Radat, F., Chanraud, S., Di Scala, G., Dousset, V., and Allard, M. (2013). Psychological and neuropsychological correlates of dependence-related behaviour in medication overuse headaches: A one year follow-up study. *The journal of headache and pain*, 14(1):59. PMID: 23826963 PMCID: PMC3706211.

Raftery, A. E. and Lewis, S. M. (1992). [practical markov chain monte carlo]: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, 7(4):493–497. ArticleType: research-article / Full publication date: Nov., 1992 / Copyright 1992 Institute of Mathematical Statistics.

Reynolds, B., Ortengren, A., Richards, J. B., and de Wit, H. (2006). Tdimensions of impulsive behaviour: Personality and behavioural measures. *Personality and Individual Differences*, 40:305–315.

Salgado, J. V., Malloy-Diniz, L. F., Campos, V. R., Abrantes, S. S. C., Fuentes, D., Bechara, A., and Correa, H. (2009). Neuropsychological assessment of impulsive behavior in abstinent alcohol-dependent subjects. *Revista brasileira de psiquiatria (So Paulo, Brazil: 1999)*, 31(1):4–9. PMID: 19506769.

Shmueli, G. (2010). To explain or predict? *Statistical Science*, 25(3):289–310.

Silva, I. R. (2017). On the correspondance between frequentist and bayesian tests. *Communications in Statistics - Theory and Methods*.

Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., and Wagen-

makers, E.-J. (2013a). Performance of healthy participants on the Iowa gambling task. *Psychological Assessment*, 25(1):180–193.

Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (2013b). A comparison of reinforcement learning models for the Iowa gambling task using parameter space partitioning. *The Journal of Problem Solving*, 5(2).

Stocco, A., Fum, D., and Napoli, A. (2009). Dissociable processes underlying decisions in the Iowa gambling task: A new integrative framework. *Behavioral and Brain Functions : BBF*, 5:1–1. PMID: 19121215 PMCID: 2645419.

Stout, J. C., Busemeyer, J. R., Lin, A., Grant, S. J., and Bonson, K. R. (2004). Cognitive modeling analysis of decision-making processes in cocaine abusers. *Psychonomic Bulletin and Review*, 11(4):742–747.

Stout, J. C., Rock, S. L., Campbell, M. C., Busemeyer, J. R., and Finn, P. R. (2005). Psychological processes underlying risky decisions in drug abusers. *Psychology of Addictive Behaviours*, 19:148–157.

Sullivan, K. J., Shadish, W. R., and Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological methods*, 20(1):26–42.

Thai, P. K., Lai, F. L., Bruno, R., van Dyken, E., Hall, W., O'Brien, J., Prichard, J., and Mueller, J. (2016). Refining the excretion factors of methadone and codeine for wastewater analysis: Combining data from pharmacokinetic and wastewater studies. *Environment international*, 94:307 – 314.

Thakur, A. K. (1991). *Model: Mechanistic vs Empirical*, pages 41–51. Springer US, Boston, MA.

The R Development Core Team (2009). *R: A Language and Environment for Statistical Computing. Version 2.10.1.*

Tryon, W. W. (2016). Underreliance on mechanistic models: Comment on ferguson (2015). *American Psychologist*, 71(6):505–506.

van Nuijs, A. L., Mougel, J. F., Tarcomnicu, I., Bervoets, L., Blust, R., Jorens, P. G., Neels, H., and Covaci, A. (2011a). Sewage epidemiology: A realtime approach to estimate the consumption of illicit drugs in brussels, belgium. *Environment International*, 37(3):612–621.

van Nuijs, A. L., Pecceu, B., Theunis, L., Dubois, N., Charlier, C., Jorens, P. G., Bervoets, L., Blust, R., Meulemans, H., Neels, H., and Covaci, A. (2009). Can cocaine use be evaluated through analysis of wastewater? A nation-wide approach conducted in belgium. *Addiction*, 104(5):734–741.

van Nuijs, A. L. N., Abdellati, K., Bervoets, L., Blust, R., Jorens, P. G., Neels, H., and Covaci, A. (2012). The stability of illicit drugs and metabolites in wastewater, an important issue for sewage epidemiology? *Journal of Hazardous Materials*, 239-240:19–23.

van Nuijs, A. L. N., Castiglioni, S., Tarcomnicu, I., Postigo, C., de Alda, M., Neels, H., Zuccato, E., Barcelo, D., and Covaci, A. (2011b). Illicit drug consumption estimations derived from wastewater analysis: A critical review. *Science of the Total Environment*, 409:3564–3577.

van Ravenzwaaij, D., Dutilh, G., and Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, 55(1):94–105.

van Wel, J. H. P., Kinyua, J., van Nuijs, A. L. N., Salvatore, S., Bramness, J. G., Covaci, A., and Van Hal, G. (2016). A comparison between

wastewater-based drug data and an illicit drug use survey in a selected community. *International Journal of Drug Policy*, 34:20 – 26.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., and Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48:28–50.

Wallsten, T. S., Pleskac, T. J., and Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological review*, 112(4):862–880. PMID: 16262471.

Wang, G., Shi, J., Chen, N., Xu, L., Li, J., Li, P., Sun, Y., and Lu, L. (2013). Effects of length of abstinence on decision-making and craving in methamphetamine abusers. *PloS one*, 8(7):e68791. PMID: 23894345 PMCID: PMC3722210.

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., and Wagenmakers, E. J. (2010). Bayesian parameter estimation in the expectancy valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, 54(1):14–27.

Yechiam, E., Busemeyer, J. R., Stout, J. C., and Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science: A Journal of the American Psychological Society / APS*, 16(12):973–978. PMID: 16313662.

Yechiam, E. and Ert, E. (2007). Evaluating the reliance on past choices in

adaptive learning models. *Journal of Mathematical Psychology*, 51(2):75–84.

Yechiam, E., Kanz, J. E., Bechara, A., Stout, J. C., Busemeyer, J. R., Altmaier, E. M. M., and Paulsen, J. S. (2008). Neurocognitive deficits related to poor decision making in people behind bars. *Psychonomic Bulletin and Review*, 15(1):44–51.

Yechiam, E., Veinott, E. S., Busemeyer, J. R., and Stout, J. C. (2007). *Advances in Clinical Cognitive Science: Formal Modeling of Processes And Symptoms*, chapter 3. American Psychological Association, 1 edition.

Zuccato, E., Castiglioni, S., Senta, I., Borsotti, A., Genetti, B., Andreotti, A., Pieretti, G., and Serpelloni, G. (2016). Population surveys compared with wastewater analysis for monitoring illicit drug consumption in italy in 2010–2014. *Drug and Alcohol Dependence*, 161:178 – 188.

Zuccato, E., Castiglioni, S., Tettamanti, M., Olandese, R., Bagnati, R., Melis, M., and Fanelli, R. (2011). Changes in illicit drug consumption patterns in 2009 detected by wastewater analysis. *Drug and Alcohol Dependence*, 118(2-3):464–469.

Zuccato, E., Chiabrando, C., Castiglioni, S., Bagnati, R., and Fanelli, R. (2008). Estimating community drug abuse by wastewater analysis. *Environmental Health Perspectives*, 116(8):1027–1032.

Zuccato, E., Chiabrando, C., Castiglioni, S., Calamari, D., Bagnati, R., Schiarea, S., and Fanelli, R. (2005). Cocaine in surface waters: A new evidence-based tool to monitor community drug abuse. *Environmental Health*, 4:14.